BeFree: a text mining system for the extraction of biomedical information from the literature

Àlex Bravo Serrano

TESI DOCTORAL UPF / 2016

DIRECTOR DE LA TESI

Dra. Laura I. Furlong

DEPARTAMENT DE CIÈNCIES EXPERIMENTALS I DE LA SALUT

Universitat
Pompeu Fabra
Barcelona

The research leading to these results has received support from Instituto de Salud Carlos III-Fondo Europeo de Desarrollo Regional (PI13/00082 and CP10/00524), the Innovative Medicines Initiative Joint Undertaking under grants agreements n° [115002] (eTOX) and n° [115191] (Open PHACTS)], resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and EFPIA companies' in kind contribution. The Research Programme on Biomedical Informatics (GRIB) is a node of the Spanish National Institute of Bioinformatics (INB).

Printing of the PhD thesis has received support from the Institut Hospital del Mar d'Investigacions Mèdiques (IMIM) (Convocatòria d'ajuts 2016 per a la finalització de tesis doctorals de la Fundació IMIM).

GRIB RESEARCH PROGRAMME ON BIOMEDICAL INFORMATICS

IMIM Parc de Salut MAR Barcelona
Institut Hospital del Mar d'Investigacions Mèdiques

Para mamá, papá e Iván.

Acknowledgements

Primero de todo, me gustaría agradecer y especialmente dedicar esta tesis a toda mi familia. Cuando pienso qué voy escribir en estos agradecimientos, me vienen muchas imágenes, recuerdos, sentimientos, emociones, mamá y papá… Ya hace mucho tiempo que os fuisteis, pero no ha habido día que no pensara en vosotros. Habéis estado muy cerca de mí en todo momento, en mis pensamientos, en mis actos y en mis sueños. Por eso, mis primeros agradecimientos son para vosotros. Por lo que me disteis, por lo que me enseñasteis, por vuestros sacrificios, por todo el apoyo que recibí y recibo de vosotros. Por todo esto y más, gracias. Os echo mucho de menos. Os quiero.

Un especial agradecimiento a todos mis abuelos, con los que viví momentos maravillosos. Especialmente, me gustaría agradecer a mi abuelo Pepe, que estuvo muy ilusionado cuando empecé mi doctorado. Por causas de la vida, Pepe nos dejó. Pepe, sé que estarías muy feliz y orgulloso de verme llegar hasta aquí. Sé que lo estás. Gracias.

Muchísimas gracias a mis tíos!!! Luisa, Ginés y Beaaa!!! Siempre, siempre y siempre me habéis apoyado en todo momento y habéis estado muy pendientes y atentos de mí y de mi trabajo. Siempre os estaré muy agradecido por ello. De verdad, muchísimas gracias!

Y por último y no menos importante, me gustaría agradecer especialmente a mi hermano, Iván. Por todo, absolutamente todo lo

que ha hecho por mí. Iván, no tengo palabras para describir lo agradecido y orgulloso que estoy por tenerte como hermano, y que tienes como compañera de vida a una persona igual de excepcional que tú. Siempre has sido y eres todo un referente para mí. Chiqui! Te quiero mucho!

Chi l'avrebbe mai detto che durante il mio dottorato avrei conosciuto una delle persone più importanti della mia vita, Eleonora. Sei diventata la mia compagna, il mio motivo, il mio sogno, il mio futuro, la mia famiglia, la mia Eleonora. Lo sai quanto sono grato di averti nella mia vita, però adesso, voglio ringraziarti per tutto l'appoggio incondizionato che mi hai sempre dato e che mi è servito per andare avanti e finire questa tesi. Dopo di me, tu sei stata quella che più ha sofferto questa fase. Mi hai ascoltato, mi hai capito e motivato. Sei la persona che più ha creduto e crede in me e nel mio lavoro. Semplicemente, grazie per come sei amore mio. Ti amo tanto. Vorrei anche ringraziare tutta la tua famiglia perché mi ha sostenuto tantissimo, specialmente la tua mamma e tua sorella Virginia.

Muchísimas gracias a todos mis amigos, que habéis estado apoyándome siempre. Sería muy complicado dedicaros unas frases a cada uno de vosotros, pero ya sabéis lo más importante. Estoy muy orgulloso y agradecido de teneros. Me habéis apoyado en muchos momentos de mi vida, y os considero parte de mi família. Muchas gracias a todos y cada uno de vosotros: Edu, Ramón, Puchi, Toni, Zurdo, Corral, Gerard, Carlos, Masters, Elton, Rodrigo y

Abuelo. A los marines Dani, Capità, Koke, Ruiman, Kimet, Cunyao, Chus y Sebas. A toda la tropa de Can Dragó y Crossfit Can Dragó! De todo corazón, gracias!

Un aspecto muy importante a tener en cuenta es, que esta tesis no la he hecho yo solo. Sin la ayuda incondicional, sin el apoyo constante y sin el trabajo duro de toda la gente que he conocido durante este periodo en el PRBB, difícilmente esta tesis podría haberse llevado a cabo. Chicos, nunca olvidéis que un trocito de esta tesis os pertenece a cada uno de vosotros!

Primero de todo, me gustaría agradecer a Laura I. Furlong la oportunidad que me ofreció para poder realizar este doctorado. Laura, has sido mi jefa, mi supervisora y mi directora de tesis, pero has sido siempre tan cercana conmigo, que siempre te he visto como una compañera más. Gracias por el apoyo constante que siempre me has dado, por ayudarme en los momentos más difíciles y sobre todo por creer en mi trabajo más que yo mismo. Muchísimas gracias por todo Laura!

També m'agradaria donar les gràcies en Ferran Sanz. Per a mi has estat i ets tota una referència. Encara que la teva agenda estigués ajustadíssima, sempre t'has implicat i m'has aportat bones idees per aplicar en el meu treball. I sobretot, sempre has considerat molt positivament el meu treball. Moltes gràcies!

Y ahora me gustaría agradecer a toda la gente con la que he compartido horas de trabajo, risas, desayunos, comidas, cafés y en general, buenos momentos.

Estamos en un mundo laboral, donde frecuentemente la gente va y viene. Por eso, me gustaría seguir mis agradecimientos a aquellos que se fueron a probar nuevas oportunidades y que han compartido muchos momentos conmigo, como a Pau Carrió y Albert Antolín, y en particular a Montse, Solène y Núria.

Montse! Moltes gràcies per tot!!! Sempre m'has ajudat moltíssim, t'has interessat pel meu treball i sobre tot, m'has recolzat molt!!! Moltes gràcies Montse!! Ah! I gràcies per involucrar-te i ajudar-me a publicar el meu primer article!

Es hablar de Solène y… Pim, pam pum! Bocadillo de apuuum!!! (que no atún, porque no rima con pum). Solène!!! Sé que estás super contenta de ver que he llegado hasta aquí. Has sido una gran compañera y amiga, hemos compartido muchos buenos momentos en el trabajo y siempre me has apoyado a mí. Merci beaucoup pour tout Solène!

Nurietaaaa!!! Que haig de dir de tu??? Ets la millor!!! Hem compartit molts moments, molts sopars, molts coffee breaks i sobretot molts riures!!! Sempre, sempre, sempre m'has recolzat, m'has animat i m'has motivat en tot. M'has ajudat molt per seguir endavant. Per tot això i moltíssimes coses més, moltes gràcies Núria!!!

En los últimos momentos de tu tesis, cuando las dificultades afloran por sí solas, es cuando te das cuenta de lo geniales que son tus compañeros de trabajo (amigos). Especialmente muchas gracias a Alba y Alexia. Sabéis que no ha sido fácil, y sin pediros nada, habéis estado en todo momento a mi lado para ayudarme.

Alba, te has involucrado muchísimo para ayudarme con esta tesis. Has dejado de hacer tus cosas, te has quedado hasta tarde e incluso me has acompañado fines de semana. Como profesional eres simplemente genial, pero aún eres más grande como persona. Me siento muy orgulloso de haber compartido esta etapa contigo. Sigue así, la gente necesita personas como tú. Muchísimas gracias por todo niñaaaa!!! También muchas gracias a tus padres por su apoyo y chuches. A Lierni por todo su apoyo que siempre me ha ofrecido y a Erika por su vital ayuda.

Santa Alexia… Siempre te estaré muy agradecido. Lo que has hecho por mí es increíble. Sinceramente, no sé qué sería de esta tesis sin ti. Sin dudarlo, te has subido a mi tren a punto de estrellarse, y lo has llevado a destino. Me has ofrecido toda tu ayuda y has dedicado muchas horas y esfuerzo a revisar mi trabajo. Nunca podré agradecer todo lo que has hecho por mí. Pero sinceramente, gracias, Ευχαριστώ πολύ για όλα Αλεξία!

Carinetaaaa, m’has recolzat moooolt! Gràcies! También a Maria y Martina! Gracias a mis ITs Alfons y Miguel por la paciencia de aguantarme. A los CRGs Rodny, Gabriel y Luís. Imma! Sempre m’has recolzat i has estat molt atenta per mi, moltes gràcies! Emilio, me has conocido en el final de mi etapa, aun así, siempre has estado ahí como si hubiéramos empezado juntos, me has animado, me has escuchado y me has apoyado, simplemente, gracias.

I would like to thank also my language exchange partner, Catherine. Thank you very much for helping me to improve my terrible English. You have been very patient with me, and the truth is that with you I have improved a lot. Большое спасибо за вашу помощь!

To conclude, I would like to thank the “Su Lab”, especially Toby, Ben and Andrew. You trusted in my work and offered me the opportunity to collaborate together to participate at the BC5 challenge. Thanks to you I learned a lot and we published a new article. Thank you so much guys!!!

Abstract

Current biomedical research needs to leverage the large amount of information reported in scientific publications. Automated text processing, commonly known as text mining, has become an indispensable tool to identify, extract, organize and analyze the relevant biomedical information from the literature.

This thesis presents the BeFree system, a text mining tool for the extraction of biomedical information to support research in the genetic basis of disease and drug toxicity. BeFree can identify entities such as genes and diseases from a vast repository of biomedical text sources. Furthermore, by exploiting shallow and deep syntactic information of text, BeFree detects relationships between genes, diseases and drugs with a performance comparable to the state-of-the-art.

As a result, BeFree has been used in various applications in the biomedical field, with the aim to provide structured biomedical information for the development of knowledge and corpora resources. Furthermore, these resources are available to the scientific community for the development of novel text mining tools.

Resum

Avui dia, la recerca biomèdica ha d'aprofitar i explotar la gran quantitat d'informació inclosa en publicacions científiques. El processament automàtic de text, habitualment conegut com mineria de text o *text mining*, és una eina essencial per tal d'identificar, extreure, organitzar i analitzar la informació biomèdica més rellevant de la literatura.

Aquesta tesi presenta el sistema BeFree, una eina de *text mining* per l'extracció d'informació biomèdica per donar suport a la recerca de les bases genètiques de les malalties i la toxicitat de fàrmacs. BeFree pot identificar gens i malalties des d'un gran repositori de text biomèdic. D'altra banda, mitjançant informació lingüística continguda al text, BeFree pot detectar relacions entre gens, malalties i fàrmacs amb uns resultats comparables a l'estat de l'art.

Com a resultat, BeFree ha sigut utilitzat en diverses aplicacions del camp biomèdic, amb l'objectiu d'oferir informació biomèdica estructurada pel desenvolupament de recursos com base de dades i *corpora*. A més, aquests recursos estan disponibles per la comunitat científica pel desenvolupament de noves eines de *text mining*.

Preface

“Stay hungry, stay foolish” is the perfect quote to summarize the fundamental pillars of my working life. By following this philosophy, it is very difficult to lose motivation, enthusiasm, creativity and willingness to learn.

Throughout my work life, I have addressed different fields of computer science: from computer graphics to bioinformatics, including development of webs, games and databases, image processing and mobile applications. Each one of these topics has offered me a valuable experience and significant knowledge in order to solve current problems from different points of view.

I got involved in the bioinformatics area through one of my first jobs, while working as programmer of medical devices for clinical analysis. After that, I grew more and more curious about the bioinformatics field. For this reason, I majored in Biomedical Engineering with a European Master’s Degree. Following my Master’s work, I started to consider several future options, one of them being a doctoral program in bioinformatics. Therefore, I joined the Integrative Biomedical Informatics (IBI) Group in order to conduct research towards a PhD degree in bioinformatics, under the direction of Dr. Laura I. Furlong.

My research has, mainly, focused on natural language processing and addressed the difficult task of detecting information from the vast repositories of the biomedical literature. Biomedical research is known to be knowledge-rich and biomedical facts and findings are being reported on millions of publications. In the undertaken PhD work, text mining methods have been used with the purpose of extracting relevant biomedical information from the literature and contributing with structured knowledge to the community. For this reason, the BeFree system has been developed and will be presented in this thesis.

This thesis is organized as follows: the challenging task of extracting information from the continuously-growing scientific literature will be introduced in Chapter 1. Furthermore, the need of text mining methods for the extraction of relevant biomedical information will be discussed and current text mining solutions in life sciences will be described. In Chapter 2, the motivation and objectives of this thesis will be presented. The BeFree system will be introduced in Chapter 3, as a text mining tool for information extraction in the biomedical field. Complementary information, applications and results of the BeFree system will be presented in Chapter 4. In Chapter 5, a discussion of the work conducted in this thesis, together with limitations and future perspectives will be provided. Conclusions will be drawn in Chapter 6. Finally, selected publications, in which this work has been applied, will be listed in the Appendix

Table of contents

List of figures

List of tables

1 Introduction

"A month in the laboratory
can save an hour in the library"
Frank Westheimer (1912-2007)

1.1 The scientific literature, an unstoppable colossus

From the earliest civilizations to the present, scientists have felt the necessity to communicate and share new discoveries, advances, findings and results to the world. Clay tablets, papyrus, letters, books and other written documents have been used by researchers as a way to record their investigations and establish the scientific literature.

The rapid growth in biological data due to the constant development and advances in life sciences, coupled with the current digital age, has caused a significant accumulation of heterogeneous and complex biomedical data (1), including biomedical text. Nowadays, scientists are still producing biomedical text as a fundamental source of knowledge and millions of facts are published every year in different types of documents, such as scientific publications, patents, theses, conference abstracts, as well as, clinical and industry reports. For this reason, the biomedical literature constitutes a rich and diverse source of biomedical information that is essential for a variety of research fields in life sciences. However, its continuous and unstoppable growth imposes a barrier to exploring such a large volume of yearly-increasing data, with more than 3,000 articles published in biomedical journals per day (2).

Fortunately, numerous bibliographic databases are, nowadays, available in the biomedical domain. One of the most important and most frequently-used repositories of scientific literature in the area of life sciences is MEDLINE (3). However, the fundamental limitation of MEDLINE is the huge amount of data that contains -comprising more than 26 million publications. A practical example is, subsequently, presented in order to better understand the amount of biomedical text that needs to be processed and the human endeavor required for the extraction of a specific type of information. Consider the case where 700,000 articles are retrieved from the MEDLINE repository, which are relevant to "biomarkers". Let us assume that an individual requires 15 minutes in order to manually process (read) an article and that he can dedicate 60 hours per week and 50 weeks per year for this task. Then, he would require more than 58 years in order to manually go over all 700,000 extracted articles. It should be also noted that, this number involves only 2.7% of the total number of publications in MEDLINE.

As it has been already pointed-out, the field of biomedical research is significantly wide. Specifically, thousands of biomolecules are being investigated around the world as potential biomarkers and the results are reported on hundreds of thousands of publications, thereby, hindering the task of performing a complete review on biomarkers. Obtaining all genes associated with a specific disease (i.e., breast cancer) is another example that demonstrates the limitations arising from manually processing the biomedical literature.

Therefore, it results imperative that alternative ways be sought to process the huge amount of information available to the biomedical research community. For this reason, automated text-processing (also known as text mining) methods are investigated in order to identify, extract, organize and analyze the relevant biomedical text in a faster, more efficient and cost-effective manner.

1.2 Text mining, the researcher’s best friend

Text mining is a field of research that has emerged due to the constantly-increasing amount of biomedical literature and the resulting need to automatically process it. Text mining methods aim at assisting the scientific community in automatically browsing the literature, as well as, in extracting, structuring and standardizing relevant biological information (4,5).

The concept of *“named entity”* and its automatic identification in a piece of text (such as, names of people, organizations and locations) was first introduced in the early 1990’s (6,7). In the mid-1990’s, the term *“text mining”* was associated with its current meaning (as defined previously). Text mining was, subsequently, accepted and established as a new discipline, resulting in an explosion of conferences, workshops, books and community challenges. By the end of the 1990’s, text mining started to be applied to the biomedical domain (8), in order to extract, structure and analyze the biomedical literature.

Nowadays, text mining tools and applications are frequently used in the biomedical domain. For example, one of most widely used text mining-solutions in life sciences is the retrieval of relevant documents by querying large bibliographic repositories or the web itself, using search engines (such as PubMed (9), Scopus (10) and Google Scholar(11)). Notably, PubMed has been established as the most popular search engine in several TM applications and is based on publications that are included in the MEDLINE database. Furthermore, various important biological databases (such as the Comparative Toxicogenomics Database (CTD) (12), the Side Effect Resource (SIDER) (13) and the

Pharmacogenomics Knowledgebase (14)) have been developed by employing text mining techniques that are directly applied to biomedical text.

As mentioned earlier, text mining describes the class of techniques that seek to automatically reveal information from available sources of scientific text, commonly termed as *natural language* text. This implies a high-complexity task, since natural language text is regarded as an unstructured type of data that has not been specifically designed for automatic processing. Furthermore, its complexity increases due to differences and continuous advances in vocabulary, terminology, language structure and style.

On the same grounds, automated processing of biomedical text is a non-trivial task, mainly due to inhomogeneities encountered in the vast number of biomedical text sources. Specifically, each type of biomedical text contains information associated with a particular field of work (e.g., medical records describe the systematic documentation of medical histories of patients, while laboratory reports describe detailed experiments) in various formats (e.g., Word, PDF or XML documents) and structures (e.g., with/o sections, abstracts, tables or images). Moreover, biomedical text can be written following different controlled vocabularies or standards, or using local spelling variations, informal languages or jargons.

Another factor of complexity is related to the different levels of accessibility of biomedical data. For example, clinical or laboratory documents are often more difficult to access than scientific literature, due to privacy and confidentially issues, as well as, due to marketing and competition in the case of industry reports and patents. On the other hand, scientific literature can be much more easily accessed. For that reason, scientific publications are the most common types of biomedical text used in text mining applications.

In summary, text mining can automatically detect and extract several types of information described in the scientific literature, thereby, providing us with structured information that would be hard to obtain otherwise, by manual processing.

1.3 Biomedical text mining

The development of text mining tools in life sciences has received a great interest in the last years due to their potential application in the extraction and discovery of biological knowledge locked in the biomedical literature (15). As shown in Figure 1.1, the number of published articles related to "text mining" experiences an exponential growth from the end of the 1990's (8).

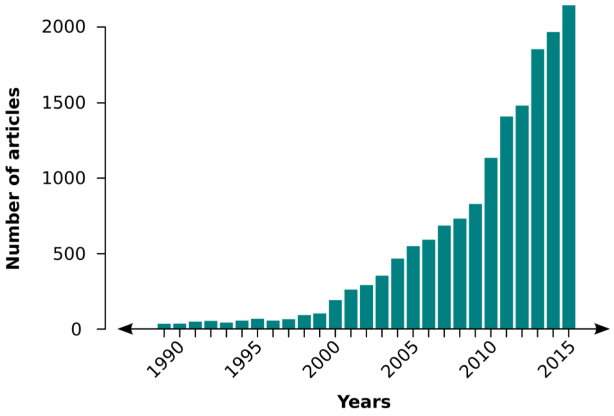


Figure 1.1. The number of articles published per year related to "Text Mining".

1.3.1 Text mining solutions for life science

Text mining techniques can assist in automatically extracting a variety of information from text sources. This includes (i) detection of biological named entities (e.g., genes, diseases, drugs, side effects, etc.), as well as, their variants and synonyms (e.g., "Neutrophil gelatinase-associated lipocalin" and "NGAL" both refer to the Lipocalin-2 gene), (ii) identification of biomedical terms by mapping them to specific entries in databases (e.g., the Lipocalin-2 gene and its synonyms mentioned above can be directly mapped to the entry "3934" in the NCBI Gene database (16) and (iii) detection of semantic relationships between concepts (e.g., genes involved in schizophrenia, biomarkers of kidney toxicity, etc.). Furthermore, text mining can be used to find answers to a variety of research questions, such as:

- Which genes cause lung cancer?
- What are the metabolites of the drug warfarin?
- Which proteins are targeted by the drug celecoxib?
- What are the protein partners of survivin protein in colon cancer cells?

In addition, during the past years, many community challenges have emerged (such as BioCreAtIvE (17–19), JNLPBA (20), CALBC (21,22), BioNLP (8,23–25) and LLL (26)) with the purpose of promoting text mining in the biomedical domain and advancing the state-of-the-art. These initiatives have fostered the development of novel tools and methods, databases, corpora, ontologies and a wide variety of semantic resources.

Currently, text mining applications can be organized in many categories according to their specific objectives: *information retrieval*, *information extraction*, *knowledge discovery*, *question-answering*, *document classification* and *document summarization*. In Figure 1.2, a schematic is shown to illustrate the different applications of text mining in life sciences. In the subsequent sections, the processes of information retrieval, information extraction and knowledge discovery will be described. Emphasis will be given on the part of information extraction, on which this thesis is mainly focused. In particular, the associated tasks of *named entity recognition (NER)* and *relation extraction (RE)* (see Figure 1.2) will be explained in more detail.

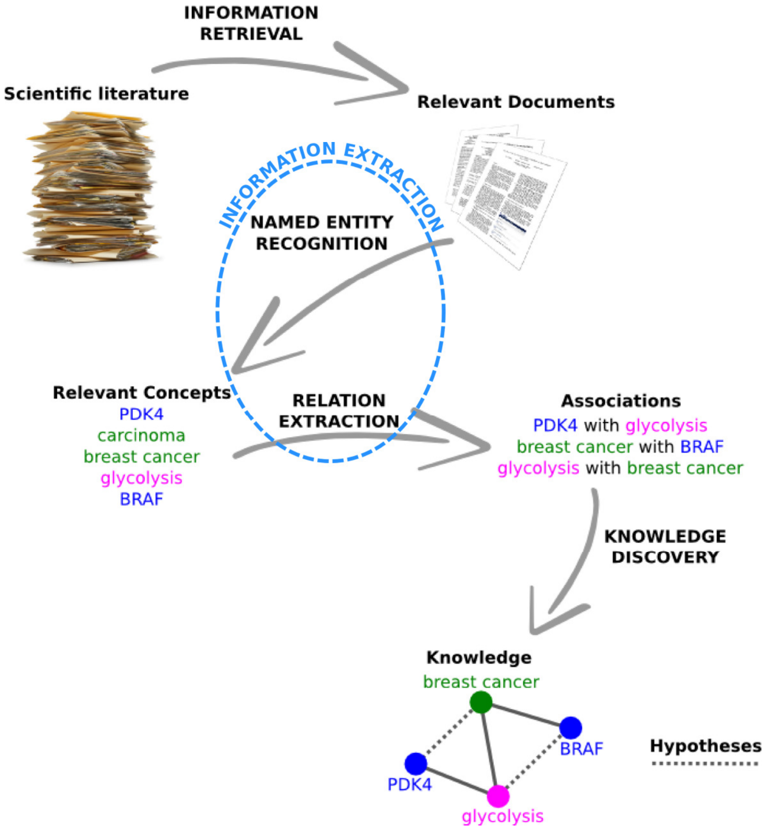


Figure 1.2. A schematic illustrating the typical text-mining workflow applied to the scientific literature.

1.3.2 Information retrieval

Usually, a text mining strategy begins with an information retrieval (IR) step (5,27). IR systems provide relevant documents extracted from a large bibliographic repository that focus on a specific topic and satisfy a certain type of input (typically introduced in the form of a query). Among publicly available IR tools, PubMed is widely used by the biomedical community (9).

In order to answer the first question in page 8 (i.e., which genes cause lung cancer?), we could use PubMed by searching for the terms "lung cancer" and "genetics". This would result in more than 22,000 citations (date of search Aug 2016). However, more advanced text mining tools are required in order to identify the individual genes that cause lung cancer from the corpus of 22,000 articles obtained.

1.3.3 Information extraction

Following the identification of relevant documents by IR tools, information extraction (IE) techniques are required to analyze the content of the retrieved documents. The IE is a crucial step in the identification, extraction and structuring of data available in the literature. The IE process is, commonly, composed of two tasks: named entity recognition (NER) and relation extraction (RE).

In named entity recognition, a NER tool detects regions of text that make reference to particular biological entities (such as "Caspase 3", "N-(2-Hydroxypropyl) methacrylamide" and "Parkinson's disease" referring to a gene, a chemical and a disease entity, respectively).

Subsequently, by applying *named entity normalization*, each biological entity can be identified with the corresponding entry in a specific biological database, ontology or terminology (e.g., the "Caspase 3" term is identified with the NCBI Gene Identifier "836" and the "Parkinson's disease" with the UMLS concept (or CUI) "C0030567" from the UMLS Metathesaurus (28)).

NER is a challenging task due to the ambiguity of the named entities in the biomedical domain. The ambiguity of entities arises in the case where certain terms used to denote a biomedical entity are also used to refer to different entities (not necessarily a biological concept nor the same biological entity type). Five biomedical entities recognized from the literature (i.e., "PDK4", "carcinoma", "breast cancer", "glucolysis" and "BRAF") were shown in Figure 1.2.

Once the entities have been detected and identified in the text, RE techniques are applied with the purpose of detecting semantic relationships between them. Typically, RE tools are aimed at identifying binary-associations between biomedical entities (e.g., protein-protein interactions and gene-disease associations). Three relationships were detected from all five biomedical entities associated with Figure 1.2 (i.e., "PDK4" with "glycolysis", "breast cancer" with "BRAF" and "glycolysis" with "breast cancer").

However, several RE systems are able to detect more complex associations (involving more than two entities) and also, to recognize direction, type and nesting of biomedical events (such as regulation, binding, gene expression and phosphorylation). BioNLP'09 Shared

Task was the first community challenge that addressed the task of *event extraction* (23).

1.3.4 Knowledge discovery

After IE processing, the biomedical information extracted from the literature can be further processed and analyzed with the purpose of data integration with other resources, as well as, the development of databases, construction of ontologies, assistance of data curation and building of semantic networks.

In recent years, knowledge discovery has emerged as a potential application after IE. IE extracts know facts (e.g., biomedical relationships) reported in the literature. Knowledge discovery focuses on the search of novel potential facts by inferring new relationships (29–35). For example, identifying genes related to diseases requires laborious experiments. Therefore, extracting candidate genes before experimental analysis could help to save time and effort (29). Knowledge discovery techniques can be applied for gene-disease associations in order to extract hidden relationships from the literature (e.g., in Figure 1.2, “glycolysis” with “BRAF” is an inferred relation) (29–31,35).

1.3.5 Metrics to evaluate text mining approaches

In the field of IE, text mining approaches are usually evaluated with a small group of related metrics. These metrics are: *precision (P)*, *recall (R)* and the *F-score (F)* (36). In particular, these metrics are generally based on *true positives (TP)*, *true negative (TN)*, *false positives (FP)* and *false negative (FN)* results.

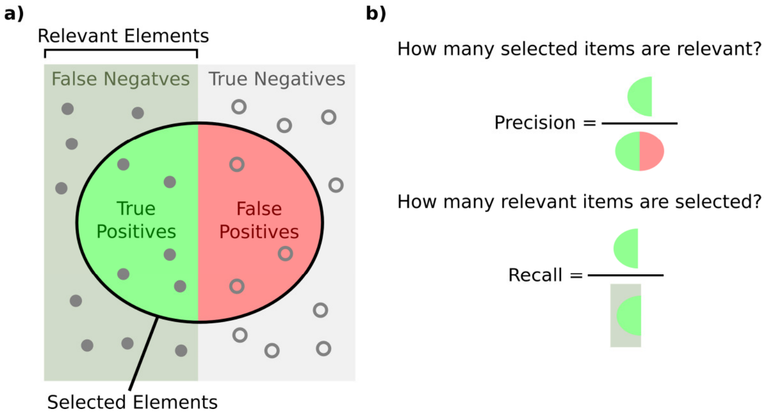


Figure 1.3. (a) Graphical description of *true positives (TP)*, *true negatives (TN)*, *false positives (FP)* and *false negatives (FN)* used in the evaluation of IE systems. (b) Graphical definition of *precision* and *recall*. Figure obtained from https://en.wikipedia.org/wiki/Precision_and_recall.

Precision measures the number of correctly identified items (*TP*) as a percentage of the number of items identified (*TP* plus *FP*). In other words, it measures how many of the items that the system identified were correct.

$$P = \frac{TP}{TP + FP}$$

Recall measures the number of correctly identified items as a percentage of the total number of correct items. To put it more simply, it measures how many of the items that should have been identified really were identified by the system.

$$R = \frac{TP}{TP + FN}$$

In simple terms, high *P* means that the system returned substantially more relevant results than irrelevant, while high *R* means that the system returned most of the relevant results. For instance, in the named gene recognition task, *P* is the total of correct gene names detected by the NER system divided by the total number of gene names detected by the NER system (correct and incorrect names). While *R* is the total of correct gene names detected by the NER system divided by the total number of gene names in the corpus.

The *F-score* (*F*) is calculated as the harmonic mean of *P* and *R* and is described by:

$$F = \frac{2PR}{P + R}$$

1.4 Information extraction in the biomedical domain

1.4.1 Extraction of relevant biomedical information

The completion of the Human Genome Project has led to a rapid increase in the number of publications in this area. This has also affected the IE domain, where the majority of text mining methods have been applied for the extraction and identification of gene/protein names and their relationships.

Information regarding diseases has risen significant interest not only within the genomics community, where the “disease” is the highest non-bibliographic information requested from PubMed (37), but also in a broader community, as justified by the increasing number of diseases tracked by using Google trends (38). It should be noted that, disease information is, typically, found to be queried together with chemical/drug or gene/protein information (39).

In this context, defining the genetic architecture of diseases and understanding disease biology appear to be key goals in the field of genome medicine. On the other hand, understanding the influence of the genetic variation of genes on drug response and drug toxicity constitute key goals in the field of pharmacogenomics. IE techniques can be applied in both research areas, with the purpose of identifying, for example, disease candidate genes and the biological impact of disease sequence variants and also (29,40), detecting genetic variations on drug response (41).

A considerable part of this thesis has focused on the area of IE with applications on genome biology and pharmacogenomics. Among the specific objectives (described in detail in Chapter 2) is the identification of human genes and their relationships with diseases, as well as, the relationships of genes and diseases with drugs.

In the following sub-sections, the origin, motivation, methodologies and current state-of-the-art of named entity recognition (NER) will be described, with emphasis on genes and diseases. Furthermore, the task of relation extraction (RE) will be presented, by focusing on relationships between genes, diseases and drugs.

1.4.2 Natural Language Processing (NLP)

Natural Language Processing (NLP) methods (15,42) are frequently used by IE approaches to extract a variety of linguistic information or features from the text, such as orthographic features (e.g., capitalizations, numbers and Greek letters), morphologic features (e.g., words, n-grams, suffixes and prefixes), syntactic features (e.g., part-of-speech (POS), phrase structure and syntactic dependencies), context features (e.g., bad-of-words, word frequencies or distances) and semantic features (e.g., target names and key-words).

NLP-based methods consist of a stack of linguistic analysis of increasing sophistication to progressively interpret language contained in text. Starting with the *tokenization* (43) to detect sentence and word boundaries and tagging the part-of-speech (e.g., noun or verb) for each word, progressing to semantic analysis for tagging of relevant entities (e.g., genes and diseases) or trigger-words (e.g., associated, activation, interaction and repression), and ending with the sentence structure (e.g., syntactic parsing) to represent the relationships (44).

Multiple tools (known as parsers) have been designed in order to extract different linguistic information from sentences. Specifically, a parser is any algorithm that converts sentences (such as a simple string of characters) into a representation that describes the linguistic information contained (e.g., a graph or tree structure).

Parsers are often used to construct the sentence structure (e.g., syntactic tree) representing syntactic relationships. According to the type of representation, parsers are commonly divided to *constituency* and *dependency parsing*. Clegg and Shepherd et al (2007) presented a wide study on dependency parsing, including constituency parsing (45).

Constituency parsing (or treebank parsers) recursively break the sentence down into clauses and phrases and produce a tree structure, where the root represents the sentence as a whole, non-leaf nodes are constituents (e.g., noun-phrase, verb-phrase and prep-phrase) and the leaves represent words. From the first binary division (a *sentence (S)* is composed of *noun-phrase (NP)* and *verb-phrase (VP)*), constituency parsing generates a one-to-one-or-more correspondence between nodes.

Clegg AB, Shepherd AJ. Benchmarking natural-language parsers for biological applications using dependency graphs. BMC bioinformatics. 2007 Jan 25;8(1):1.

Constituency Tree

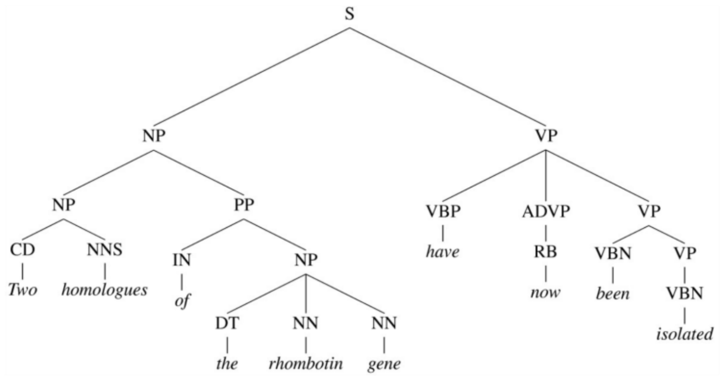
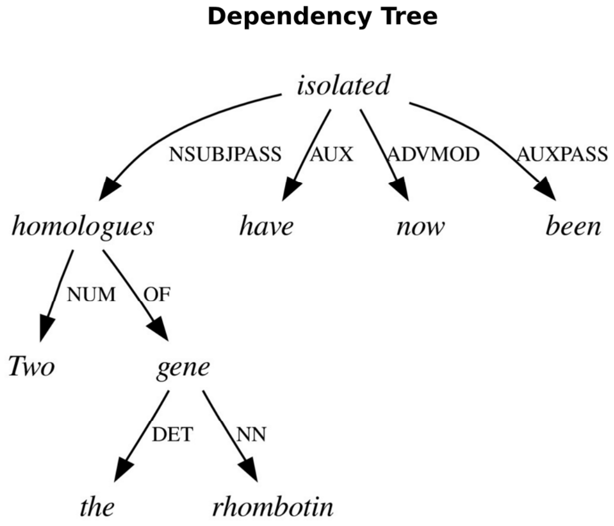


Figure 1.4. The constituency tree structure of the sentence *'Two homologues of the rhombotin gene have now been isolated'* from PMID:2034676.

In contrast, dependency parsing builds a different type of tree structure, where each node represents one word in a sentence and they are one-to-one connected by syntactic dependencies. Additionally, edges in the tree are labeled by the relationship or syntactic dependency between words (e.g., *noun singular (NN)* and *adverbial modifier (ADVMOD)*).



Clegg AB, Shepherd AJ. Benchmarking natural-language parsers for biological applications using dependency graphs. BMC bioinformatics. 2007 Jan 25;8(1):1.

Dependency Tree

Figure 1.5. The dependency tree structure associated with the sentence of Figure 1.4.

Syntactic parsing can also be classified in *shallow* (or partial) and *deep parsing*. Shallow parsing (also known as *chunking*) typically identifies noun, verb, preposition phrases, and so forth in a sentence, while deep *parsing* builds complete trees representing a sentence.

Some parsing tools have been developed and optimized for biomedical text. A majority of IE applications use the shallow parsing tools, such as Penn Treebank Tag Set (46), GENIA Tagger (47), Illinois Shallow Parser (48) and Apache OpenNLP (49) (note that not all shallow parsers identify the same type of phrases). However, the use of deep parsing techniques is gaining interest in biology applications. Several studies have reported the superiority of such techniques in extracting information from biomedical text, since they take into account the manner in which relations are represented in the text (50–52). Deep parsing tools include the Stanford Lexical Parser (53), Link Parser (54), Enju Parser (55), Genia Dependency Parser (56) and Bikel Parser (57), among others.

1.4.3 Named entity recognition (NER)

1.4.3.1 Identifying biomedical entities in text

The NER task allows the detection and identification of biomedical named entities in text. A *named entity* or *mention* is defined as a name or term used in a source of text that represents a specific entity, in this case, a biomedical entity (e.g., in the sentence '*These results are discussed in relation to neuroprotection and toxicity of the age-related pathology of AD*' from PMID: 11391700, the mention "AD" represents the Alzheimer's Disease entity or concept) (58). A biomedical entity (such as genes, proteins, drugs, tissues, diseases, mutations, pathways, species and chemicals, etc.) can be represented in the literature by several named entities or mentions.

The major problems encountered in the NER task are the complexities, inconsistences, synonyms and ambiguities -associated with the biomedical vocabulary- that make the task of detecting biomedical entities very challenging. These problems are described below.

First, the field of biomedicine is characterized by complex naming conventions and specialized terminology. For instance, 85% of the names in the biomedical field consists of more than one word (59) and frequent use of ad-hoc abbreviations (such as, the "TRADD" and "TRAF2" abbreviations found in the following piece of text '*Moreover, NF-κB activation induced by overexpression of the TNF* *receptor–associated proteins, TNF receptor–associated death domain protein (TRADD), receptor interacting protein (RIP), and TNF receptor–associated factor 2 (TRAF2) was also inhibited by expression of A20…*' extracted from (60)).

Second, the terminology used in the biomedical literature is often inconsistent. Chen et al., 2005 (61) reported that 75% of the gene mentions found in the literature did not follow established conventions (as official symbols or full names). Also, the terminology suffers from different spelling variations, such as morphological (e.g., "tumor" vs. "tumour" and "anemia" vs. "anaemia"), orthographic (e.g., "NF-kB", "NFkappaB", "NF-kappa-B" are some variations of the "NF-kappa beta" term) and composed variations (e.g., "cardiac and respiratory complications" refer to two concepts, "cardiac complication" and "respiratory complication").

Third, the biomedical entities comprise a large amount of terminology including a lot of synonyms constantly evolving. For example, "Neutrophil gelatinase-associated lipocalin", "Oncogene 24p3", "NGAL" and "LCN2" are synonyms referring to the Lipocalin-2 gene; or "Hepatolenticular degeneration", "Progressive lenticular degeneration" are synonyms related to Wilson's disease.

Finally, the most controversial issue affecting the NER task is the high ambiguity involved in the biomedical terminology. Ambiguity is a linguistic phenomenon of a term, which can represent different meanings. As seen in the previous example, the ambiguity may cover distinct cases:

(1) Some entities contain terminology that coincides with common English words (e.g., “can” refers to a verb or a synonym of the NUP214 gene).

(2) Different biomedical entity types may share terminology (e.g., the “PSA” term can refer to “Prostate-specific antigen” protein, “Puromycin-sensitive aminopeptidase” protein, “Protein S alpha” gene and “Psoriatic arthritis” disease). This issue is particularly tackled in this thesis, in order to disambiguate mentions between genes and diseases that may share a potential overlap in their terminology.

(3) Furthermore, several terms referring to biomedical entities may coincide with terms with a completely different meaning (e.g., the “PSA” term can refer to “Pharmaceutical Society of Australia” or “Political Studies Association”).

(4) In addition, new biomedical entities are constantly being discovered and assigned to names, which could be already in use or coincide with common English words.

With regard to the entity normalization process, ambiguity arises when a mention is linked to multiple identifiers, that is, several biomedical entities of the same type may share terminology. In the previous example, if “PSA” was classified as a protein, it could be mapped to “P07288” (“Prostate-specific antigen”) and “P55786” (“Puromycin-sensitive aminopeptidase”) entries from the UniProt database (62).

1.4.3.2 NER approaches

During the past years, several NER systems have been developed in order to identify one or multiple biological entities in the scientific literature. The main approaches used for the detection of biological entities can be divided into three categories: *dictionary-based* approach*,* *rule-based* approach and *supervised learning* approach (63).

The dictionary-based approach is the most common technique in the biomedical NER domain. Dictionaries are large collections of terms representing biomedical entities. Different kinds of algorithms are applied to look-up matches between a piece of text and a term from the dictionary.

The major advantages of this approach are that it is straightforward and facilitates the entity normalization, by linking each term to the corresponding database identifier. The entity normalization process is a crucial requirement in the majority of text mining applications in the biomedical domain. Specifically, it permits the integration of the information extracted in the literature with related knowledge from other biomedical resources using standard identifiers. However, dictionary-based approaches are limited to detecting only terms that are included in the dictionaries (64).

The rule-approach consists of a set of decision rules previously structured that satisfy each biological entity. These rules describe naming structures for different entities by using linguistic information (such as orthographical, morphological and syntactic information, for more detail see Section 1.4.2). Rules can be manually defined (65), although several algorithms have been developed in order to automatically obtain rules from text or terminology (66,67).

Decision rules have greater freedom than dictionaries, in order to detect unseen or newly-discovered entities (e.g., the detection of protein by noun-phrases including the final word "receptor"). In contrast, decision rules are usually based on specific corpora and are often not effective in all cases (64).

The supervised learning approach uses annotated data or corpora (see Section 1.5 for more detail) to "learn" useful information (e.g., linguistic features) in order to subsequently detect mentions in text. Supervised learning approaches learn linguistic features involving entities, that is, the manner in which language represents biomedical entities in the literature. So, these approaches have a greater flexibility to detect unseen or newly-discovered entities. In contrast, they require correctly labeled corpora, thereby, requiring an enormous effort to build them, frequently by manual annotation (21).

From the earliest learning-based text mining systems to the present day, different supervised approaches have been developed in order to detect biomedical entities: Support Vector Machines (SVMs) (68,69), Hidden Markov Models (HMMs) (70,71), Maximum Entropy (ME) (72,73), Naïve Bayes (67), Conditional Random Fields (CRFs) (64,74) and others learning approaches (75–77).

Hybrid approaches are often used to take advantage of different techniques (64,67). Specifically, dictionary-based are often combined with rule-based approaches (78) or machine learning approaches (79), because only dictionaries can resolve the entity normalization process between named entities and their identification.

1.4.3.3 Named gene recognition (NGR)

The Named Gene Recognition (NGR) process involves the *Gene Mention (GM)* task for the detection of gene and protein mentions (or names) and the *Gene Normalization (GN)* task for the normalization of mentions with the corresponding unique identifier.

In late 1990's, the first works on biomedical text mining were focused on the detection of gene/protein names in the biomedical literature (66,80,81). Until nowadays, new methodologies and strategies are being developed with the purpose of improving the detection of genes.

Although several efforts were, initially, proposed to establish and promote the use of a standard nomenclature for genes (such as the *'Guidelines for Formatting Gene and Protein Names'* (82), or the guidelines for human gene nomenclature from the *Human Gene Nomenclature Committee (HGNC)* (83)), only 25% of the gene names mentioned in the literature follow these conventions (as official symbols or full names) (61). This 'creative' approach followed by the authors has resulted in a wide variety of gene terminology without clear rules concerning gene nomenclature, thereby, exacerbating the ambiguity issue.

In addition, gene and protein entities often share similar terminology and also, nomenclature guidelines are not widely adopted by authors, such that, a strict distinction between them is not commonly made, thereby, leading to the use of gene and protein names interchangeably. This, in turn, may cause problems in distinguishing the two entities in the literature. Therefore, it results difficult for a reader, as well as, for an automatic system, to determine whether the specific mention refers to a gene or to its corresponding protein. Consequently, most of NGR systems detect gene/proteins as one single biomedical entity.

Significant ambiguity is encountered in the gene names associated with different animal species (known as orthologous genes). This affects mainly the GN task. For example, the Interleukin 6 gene can be linked to more than 150 entries in the NCBI Gene database depending on the species under study (e.g., the entries “3569”, “399500” and “403985” correspond to human, pig and dog species, respectively).

The first NER strategies and systems that extracted gene and protein names from the literature were focused only on the GM task (such as PROPER (66), Yapex (84), AbGene (67), GAPSCORE (69) and BANNER (85)).

It was not until 2004, that the BioCreAtIvE (17) challenge started to address the identification of gene/protein names in text. BioCreAtIvE proposed a NGR challenge, involving the GM task (86) and the GN task (87). Both tasks were focused on three specific species: yeast, fly and mouse. In summary, 4 teams achieved an *F-score* (see Section 1.3.5) greater than 80% with respect to the GM task, while the results obtained for the GN task were also high (*F-scores* of 92% , 82% and 79% were achieved, respectively, for yeast, fly and mouse).

In 2007, BioCreAtIvE II (18) was organized with a GN task focused on human genes (88). Furthermore, it included a GM task that achieved a total *F-score* of 87%. The GN task was an important reference point in the detection and identification of human genes in the literature. Twenty teams participated in the GN challenge. In most cases, the teams adapted their developed and evaluated systems for the GM task or other well-known systems (such as BANNER (85) and ABNER (89)), including additional processing steps to address the GN task.

The best performance in the GN task was an *F-score* of 81% achieved by Hakenberg et al. (2007) (90), while the median result between all twenty participating teams was 73%. In their work, Hakenberg et al. (2007) (90) enriched the provided lexical resource with additional synonyms from NCBI Gene. The terminology was processed to generate new variations and fixed rules. The rules were matched against the text in order to extract gene mentions and each mention was assigned different candidate identifiers. Subsequently, the erroneous results (or *false positives*, see Section 1.3.5) were eliminated using an alternative score (in this case, based on the frequency with which the term appears) and regular expressions to detect mentions referring to tissues, cell lines, molecules, etc. Finally, for the identifier disambiguation task, they used a scoring approach. The context of the identified mention in the abstract was compared to the contextual information of the gene obtained from different resources (such as Go annotations, functions, locations, mutations, GeneRIFs, UniProt keywords and diseases).

BioCreAtIvE promoted new text mining systems and strategies in the *NGR* domain, contributing to advances of the state-of-the-art. These systems and strategies established the foundation of the detection and identification of any biomedical entity and are based, mainly, on four steps:

(1) Detecting regions in the text (GM task) as mentions of gene and protein names. This was achieved using different approaches, such as, building specific rules (90), developing a dictionary of terms in order to find matches against the text, combining exact and approximate matching (91), (92), or using machine learning approaches in the detection process (93).

(2) Applying a mapping approach to link detected gene mentions to a gene identifier list.

(3) Developing strategies to deal with the ambiguity problem between gene identifiers (known as Word Sense Disambiguation (WSD) strategies). Different approaches were applied in this step, such as scoring approach (90), alternative dictionaries (91) and context-based approach (92).

(4) Applying a post-filtering process in order to remove *false positive* results reported by the *NGR* process. In the majority of the presented works, the post-filtering step is based on

different/various scoring methods (90,94), or rule-based approaches (91).

After BioCreAtIvE challenges (including BioCreAtIvE III with a new GN challenge (95)), various works were developed to address the entire *NGR* task. The great majority of these works were evaluated based on BioCreAtIvE II GN corpus, as it will be described below.

Hakenberg et al. (2008) presented the GNAT system, an updated version of the winner approach in the BioCreAtIvE II GN task. They extended the system with a machine learning approach, specifically the CRF approach , 85.4% *F-score* (93), by combining machine learning (using orthographic, morphologic and shallow parsing features), rules and dictionaries.

Wermter et al. (2009) developed GeNo, a high-performing system for gene name identification. Their system is based on scoring approaches to resolve ambiguous gene names. Moreover, their system was evaluated against the BioCreAtIvE II GN corpus, obtaining 86.4% of *F-score*.

Lately, Hu et al. (2012) achieved a performance of 83.5% of *F-score*. Although this was lower than the previously reported scores, their approach was developed based on their GM system that had achieved the highest performance in the BioCreAtIvE II GM task (89% of *F-score*).

Recently, two important systems have been presented. Li et al. (2013) (96) reported the highest performance on the BioCreAtIvE II GN corpus (90.1%. of *F-score*) and Wei et al. (2015) (97) presented one of the best performing systems for gene name disambiguation across species. In particular, Li et al. (2013) described a creative combination of approaches to achieve the best result in the BioCreAtIvE II GM task. They, initially, applied this previously developed gene mention tagger (98) to detect gene names. The normalization step was based on dictionary matching. They combined exact-matching with soft-matching to get candidate identifiers for each gene name. Then, a similarity semantic algorithm was applied to select the correct identifier. Finally, a filter step based on Wikipedia knowledge was applied to remove *false positives*.

Wei et al. (2015) implemented a supervised learning approach based on CRF. Their CRF-based approach was trained with shallow and context linguistic information to detect gene names in text. For the gene name disambiguation, several rules were applied to the CRF results (e.g., for long form-abbreviation pairs, the features of the long form mention are prioritized) to remove erroneous gene names. For the normalization step, they applied their previous tool (99,100), which was based on a statistical inference network model. In the human gene normalization task (BioCreAtIvE II GN), an *F-score* of 86.7% was obtained. Additionally, their system was developed to normalize genes across species. In this context, their system achieved 50.1% of *F-score* in the cross species gene normalization task (BioCreAtIvE III GN), demonstrating that gene name disambiguation across species is still an unsolved problem.

1.4.3.4 Named disease recognition (NDR)

Compared to the identification of genes, the identification of diseases has received less attention in the text mining community (101,102). In comparison with genomic domain, a lower number of disease resources, providing rich information, are available.

The named disease recognition task shares many issues and challenges reported in the named gene recognition task. In contrast, the disease entity is a diffuse biomedical category, which involves and represents a wide range of medical aspects (such as diseases, disorders, symptoms, sings, treatments and adverse effects). For example, works on disease recognition may be focused only on particular concepts as drug adverse effect (103) or considering all diseases (104). This situation is also reflected in terminological resources available for diseases, which can focus on a specific disease category or domain (such as cancer or clinical domain), some of which are described below:

(1) The *MeSH (Medical Subject Headings)* is a controlled vocabulary used for indexing articles in PubMed and provides a wide branch for disease terminology (105).

(2) The *Systematized Nomenclature of Medicine – Clinical Terms (Snomed CT)* is a comprehensive and precise clinical health terminology (106).

(3) The *International Classification of Diseases (ICD)* is the medical coding and classification standard to gather information about different health conditions. ICD defines diseases, injuries,

sign and symptoms and further related health circumstances (107).

(4) The *OMIM (Online Mendelian Inheritance in Man)* database is a manually-curated database, based on diseases information (including terminology) and their involved genes (108).

(5) The *UMLS (Unified Medical Language System)* is a broad terminological resource that integrates several biomedical vocabularies and ontologies in a single source (28).

Several tools have been developed for the recognition of diseases in the clinical domain driven by community challenges, such as the i2b2 Challenge (109), ShARe/CLEF eHealth (110) and SemEval task 7 (111). However, there are significant differences between the clinical text and biomedical literature (112).

Biomedical literature is composed of books, publications, abstracts, posters, etc. In contrast, clinical text is written by clinicians in the clinical setting (describing for example, patients, their pathologies and their personal, social, and medical histories). Meystre et al. (2008) (112) reported all languages differences found in clinical text, pointing out divergent characteristics such as misspellings, ungrammaticality, the use of shorthand, informal templates and diversity of input sources (113).

TM-solutions have been developed for the *disease mention (DM)* and *disease normalization (DN)* of disease names in the biomedical literature. Currently, a large number of studies focused on the DM task can be encountered (such as (114), (85), (115), (89), (103), (116), (117), (102), (118) and (119)), while much fewer on the DN task (such as (116) and (104)).

In the DM domain, the initial studies were based on gene name recognition systems, which were adapted for the detection of disease names in the literature (e.g., ABNER, Lingpipe and BANNER with obtained *F-scores* of 53.44% (89), 51.15% (115) and 54.84% (85), respectively, on the BioText corpus). Later, Leaman et al. 2009 improved the BANNER performance, by developing a new BANNER version to detect disease names and achieved 79.9% of *F-score* on ADZC corpus (114).

In the same ADZC corpus, Chowdhury et al. (2010) showed a machine learning approach (CRF), combining a full set of linguistic information (such as orthographic and syntactic dependency features). Their results with respect to the DM task, exhibited a slight improvement with 81% of *F-score* (117).

Gurulingappa et al. (2010) adapted ProMiner with dictionaries from different terminological resources (such as MeSH, ICD-10 and SNOMED-CT) to identify diseases and adverse effects in biomedical literature (103). They made use of their own corpus to evaluate the performance and reported higher *F-scores* in disease and adverse effect matching (i.e., 80% and 71%, respectively).

Doğan & Lu (2012) evaluated BANNER in their newly developed NCBI disease corpus and achieved a higher performance (84% of *F-score*) in comparison with ADZC corpus.

Recently, Huang et al. 2013 improved the results previously obtained on BioText corpus. They proposed a machine learning approach based on CRF, trained with linguistic features extracted by non-deep parsing. They reported a slight improvement with 56.67% of *F-score*.

On the other hand, fewer works have been presented including DN. Kang et al. (2012) developed a rule-based NLP approach to improve disease normalization systems (116). They compared the performance of two known biomedical normalization systems, MetaMap (120) and Peregrine (121), applying their additional rule-based NLP approach. The evaluation of DM and DN tasks were carried out against the AZDC corpus. The results showed a significant improvement in the system performance when the rule-based NLP approach was also applied. Specifically, in the DM task, Metamap and Peregrine obtained 73% and 78% of *F-scores*. In the DN task, the *F-scores* were calculated to be equal to 66.2% and 69.8%, respectively.

More recently, Leaman et al. (2013) presented DNorm system for disease name normalization (104). DNorm was evaluated together with other approaches using a subset of the NCBI Disease corpus (122). Based on BANNER, DNorm applied a machine-learning approach (for DN) in order to compute similarities between mentions and concept names, and finally, achieved the highest *F-score* (80.9%).

1.4.4 Relation Extraction (RE)

Once the entities have been identified in the text, the following step of a text mining method is to identify relationships (or associations) between entities by using RE tools.

1.4.4.1 Identifying relationships between biomedical entities

In the context of biomedical literature, RE tools allow us to identify relationships between biomedical entities. In its simplest form, an association is defined between two biomedical entities or a binary relationship (such as drug-disease and gene-disease associations). *RE* can classify associations in more detail. For example, drug-disease associations can, further, include drug indications (123) and side-effects (124–126). Furthermore, as commented previously, associations can involve more than two entities and are usually characterized by specific actions or events that may represent fundamental molecular processes (e.g., binding and phosphorylation).

In comparison with the NER task, RE shares many common challenges and motivations, such as the creation of high quality annotated corpora, with the purpose of training and evaluating systems. It should be noted that, many works on RE have been widely applied on the genetic/proteomic field. This interest has been also promoted through community challenges (such as the LLL in the detection of genic interactions (127) or BioCreAtIvE II in the protein-protein interaction (PPI) task (128)).

In contrast, the detection of relationships is significantly more difficult than the detection of biological entities, because relationships are generally expressed in a discontinuous way (129). Specifically, a relation can be expressed in a single sentence or at the document level (e.g., spanning multiple sentences). Although recently, several works have focused on relations that span multiple sentences (for example, the BioCreAtIvE V designed a specific task for chemical-induced disease (CID) relations at the abstract-level (130)), most RE systems are, typically, based on relations expressed in a single sentence.

An overview of RE approaches will be performed in the subsequent sub-sections and emphasis will be given on the extraction of binary associations between genes, diseases and drugs.

1.4.4.2 RE approaches

Several strategies, techniques and methodologies have been described to address the problem of extracting relationships between entities from the biological literature. From systems based only on simple co-occurrences to more complex systems using linguistic analysis, RE approaches can be commonly organized in three classes: *co-occurrence-based statistics*, *pattern-based* and *supervised learning approaches* (131). However, several authors classify the RE strategies in co-occurrence-based and *NLP-based approaches*, since pattern and learning approaches are, mainly, developed by NLP-based methods (44,132).

Co-occurrence-based statistics constitute the simplest approach to detect relationships between biomedical entities (33,133,134). It assumes that two entities are associated if they are mentioned together in the same piece of text (typically a sentence). As it can be deduced, not all couples of entities mentioned together are related. For example, Chun et al. (2006) reported that only 30% of protein pairs co-occurring in the same sentences have an actual interaction (40). Then, statistical methods are applied to rank co-occurrences and to select the best potential relationships, mainly based on frequency scoring (31,135,136).

However, co-occurrence methods often report a poor *precision* (generating many *false positives*), are unable to detect the direction of the relationship and also, exhibit difficulties in the distinction between direct and indirect associations (44,137–139). The fundamental reason causing these limitations is the fact that co-occurrence does not take into account any linguistic information regarding the relationship. On the other hand, co-occurrence approaches are straightforward and thus, it is easier to obtain candidate relationships between entities without the need of more complex linguistic analysis. For this reason, most of the RE approaches are applied from co-occurrences (hybrid approaches) (40,140,141).

As it was seen in the previous section, NLP-based methods extract linguistic information, which is exploited by NLP-based approaches in order to extract specific structures representing relationships (e.g., formal grammars that specify relationship or entities connected by syntactic dependencies) (142,143).

Linguistic analysis/information is frequently used in pattern-based and supervised learning approaches. Principally, pattern-based approaches work with patterns constructed from the linguistically annotated text, which are matched against unseen text to detect relationships (note that the unseen text should be processed by NLP-methods to extract the necessary linguistic information). The patterns can be defined manually (144–148) or can be automatically constructed from annotated corpora (127,149–151).

Systems based on manually developed patterns, frequently, use specific rules and can achieve high *precision*. On the other side, many relationships are missed when compared with systems that are based on automatically constructed patterns (146). However, both approaches require a laborious manual effort to build patterns or to annotate a large corpus (44). Nevertheless, pattern forms can be too rigid to capture semantic/syntactic paraphrases or long-range relations (152,153).

Supervised learning approaches also use annotated corpora to learn linguistic information representing relationships, in order to detect new relationships from biomedical literature (30,154,155). However, availability of annotated corpora is the major limitation of such approaches and consequently, they are often significantly tailored to the specific corpus used for development. They have reported good performance in cross-validation evaluations, but the performance is often reduced when detection is scaled to other corpora (129,156).

During the last decade, supervised learning approaches have been widely presented as an appropriate alternative for relation extraction, reporting a better performance compared to simple co-occurrence-based approaches (23).

It should be noted that, most of the supervised learning approaches previously described are based on kernel methods (KMs) (141,144,152,153,157–164), conditional random fields (CRFs) (30) and maximum entropy (ME) (165). Specifically, KM have been proposed as the most popular learning approaches for relation extraction, being the nearest neighbor classification (NNC) and especially support vector machines (SVMs) the most popular examples (164).

1.4.4.3 Relations in the genome medicine and pharmacogenomics areas

Most of the efforts in the task of relation extraction have been devoted, so far, to the identification of interactions between proteins (PPIs). Besides the scientific interest, this can be also explained by the availability of corpora and the push driven by specific text mining challenges (5). In contrast, less attention has been paid to the identification of relationships between entities of biomedical interest such as diseases, drugs, genes and their sequence variants.

In the past years, this trend has, favorably, changed and there is much more interest shown in extracting such type of information (166). The first RE systems for non-PPI relationships were based on strategies or tools, originally, developed for PPIs extraction.

In particular, Giuliano et al. (2006) presented a machine learning approach using kernel methods based solely on shallow linguistic information for PPI extraction (157). Subsequently, this work was adapted to extract drug interactions (160) and drug-side effects (126). Buyko et al. (2012) also adapted the JReX system, originally developed for PPI extraction, for the extraction of multiple biomedical binary-relationships between genes, drugs and diseases (167).

There are examples of systems developed for identification of drug targets (168,169), interactions between drugs (160,170–172), drug indications (123,173), drug adverse effects (126,174), gene-disease associations (29,30,40) and also, others covering different types of relationships (41,167). Furthermore, alternative text mining systems have been presented that detect connections between triplets of entities (e.g., chemical-protein/gene-disease connections (175,176)).

Although machine learning approaches exhibit good performance when shallow linguistic parsing (126,160) is used, there have been works reporting a superior RE performance when dependency parsing (or syntactic parsing) is applied -especially when syntactic dependencies are combined with shallow linguistic information (177–179). Both supervised learning and rule-based approaches have demonstrated good performance by exploiting both shallow and dependency information (23,144).

Regarding the identification of associations between diseases and genes from the literature, a growing interest has been observed during the recent years, specifically for human diseases (29,30,40,41,180). In comparison with PPIs, drug-disease associations (e.g., drug indications and adverse drug events) or drug interactions, there are still a few text mining works aimed at extracting genetic disease information (where many of them study a specific disease such as cancer (29,40,181)). Other approaches in this area use data mining to discover diseases associated with genes, by combining data extracted from different databases (35,181) and applying simple text mining approaches.

Co-occurrences-based statistics approaches were the first text mining techniques applied in order to extract gene-disease associations (40,138,182–184). Although various text mining techniques based on co-occurrences statistics have been, so far, presented (136), most of the current systems use NLP techniques and machine learning approaches (30,40).

Chun et al. (2006) presented an early work to extract relevant genes associated with prostate cancer from MEDLINE abstracts, based on co-occurrences (40). In order to improve their results, a maximum-entropy approach was applied in order to remove erroneous gene mentions, using shallow linguistics and syntactic-dependency features. Their system was evaluated using a dataset composed of 1,000 sentences and an *F-score* of 82.5% was obtained.

Later, Chun and collages extended their work in order to classify the prostate-cancer-gene relationships in six categories (such as genetic variation, gene expression and clinical marker) (185). Specifically, they developed a machine learning-based topic-classified relation, based on

ME. They classified each relationship category with an *F-score* between 55% and 75%.

Bundschus et al. (2008) proposed a machine learning approach to extract gene-disease associations with the purpose of developing a gene-disease network from the GeneRIF database (30). Specifically, their work was centered in the extraction of GDAs and the characterization of relationships between genes and diseases (such as altered expression, genetic variation and regulatory modification). The learning approach that they proposed, was based on CRF and only simple features derived from words were used (no higher level linguistic information, such as POS tags or phrase structures, was used). By 10-fold cross-validation on a subset from GeneRIF sentences, they obtained a performance of *F-scores* of 78% in the detection of genes, diseases and their relationships and 78% in the semantic classification task.

Özgür et al. (2008) described a combined approach based on text mining techniques (with syntactic parsing) and network analysis methods to extract genes associated with breast cancer from the literature (29). They evaluated their system from a list of fifteen known genes related to prostate cancer. Initially, a set of abstracts were mined with Genia Tagger to detect genes in sentences. Subsequently, only sentences with gene co-occurrences and specific key-words (such as "coactive" and "localize") were taken. Finally, a kernel-based approach was trained with dependency information between candidate genes. Once the network was constructed, different metrics were used to evaluate the centrality of genes related to prostate cancer in the network. They demonstrated a high *precision* of 90% for the most central ten genes and 80% of *precision* for the top twenty genes.

Hakenberg et al. 2012 presented a system for extracting twelve types of binary relationships (such as gene-drug, gene-disease, gene-variant and drug-disease) based on co-occurrence between them (41). Their system was evaluated on their own corpus and demonstrated a performance of 76%, 84% and 83% for drug-disease, gene-disease and drug-target associations, respectively.

Also, Buyko et al. (2012) adapted their previously developed system (186) with the purpose of extracting gene-disease, gene-drug and drug-disease (167). They compared two machine-learning-based approaches for relation extraction: a feature-based approach (ME) with shallow linguistic information and a kernel-based approach (SVM) with dependency information. They demonstrated superior performance when the feature-based approach was used (reporting total *F-scores* approximately equal to 80%).

Hou et al. (2013) described an automatic rule extraction for gene-disease relationship extraction from a set of MEDLINE abstracts (2,000 sentences). Specifically, they built rules from 2,000 sentences and evaluated them on a test dataset composed of 400 sentences. An *F-score* of 66.9% was achieved (180).

Most recently, Pletscher-Frankild et al. (2015) presented a text mining application to extract and integrate gene-disease associations (136). They, initially, used dictionaries to detect gene and disease mentions in sentences and subsequently, the extracted co-occurrences were ranked by the scoring scheme previously presented in (187).

Xu et al. (2016) developed a complete framework to extract GDAs from literature. In the RE step, they used a SVM approach to train a model by combining lexical features and syntactic features. Subsequently, GDAs were ranked by co-occurrence frequency, paper citations and author information. They achieved a performance of 86% using their own corpus (188).

1.5 Corpora for text mining

In the context of biomedical text mining, a corpus is defined as a set of annotated documents, that is, a set of documents enriched with labeled biomedical information. Such annotated documents, usually, focus on a specific topic and can appear in the form of a full text (189,190), an abstract (191,192) or a sentence (193). Corpora can contain one or various types of annotations, usually biomedical entity names (e.g., genes (192), diseases (114) or chemicals (194)) and relationships between entities (e.g., drug-drug interactions (195), chemical-induce diseases (130), etc.).

1.5.1 Annotated corpora in the biomedical domain

The availability of annotated corpora is a basic requirement for developing new methods in text mining, in particular for supervised learning and pattern-based approaches. Furthermore, annotated corpora are frequently used to evaluate the performance of text mining systems (usually in values of *precision (P)*, *recall (R)* and *F-score (F)*, see

Section 1.3.5), thereby, allowing the comparison between different text mining systems that address a particular task.

Corpora are usually developed by manual curation, mainly by experts in a particular domain. In the TM area, a “Gold Standard” is a corpus developed by one or more human experts. Alternatively, if annotations are automatically derived, the corpus is defined as a “Silver Standard” (22).

The development of a Gold Standard Corpus requires tremendous human efforts; consequently, this task should be supported with different resources with the purpose of guiding and assisting the manual annotation (196). Mainly these resources are:

(1) Guidelines describing all biomedical and technical details involving the particular annotation task.

(2) An annotation tool is an essential component to facilitate the annotation task. The annotation tool should be user-friendly in order to focus the attention of the annotators on the particular annotation task.

(3) A training stage for annotators following the guidelines and using the annotation tool to prepare them for the annotation task, and most importantly, to obtain feedback and to understand possible discrepancies between annotators.

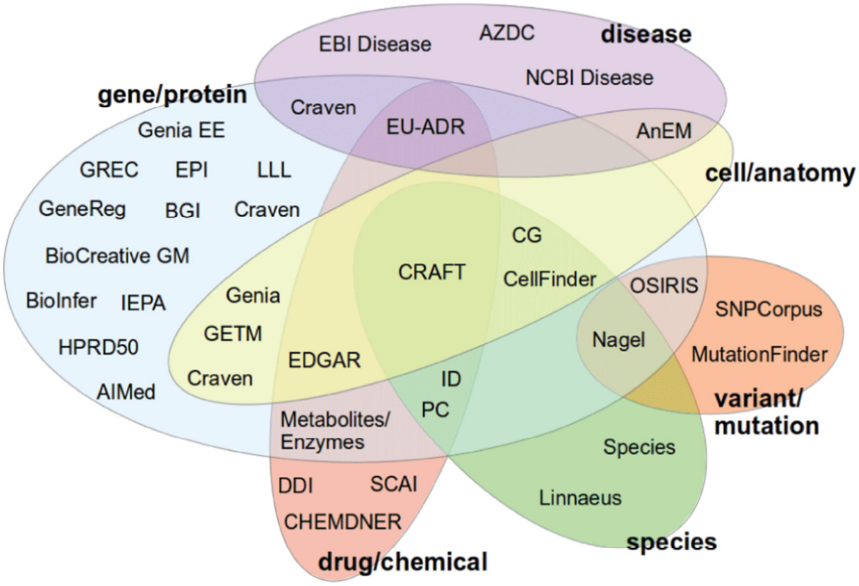
(4) An inter-annotator agreement value in order to assess the consensus achieved between annotators.

Text mining techniques are usually applied providing automatic annotations before the manual annotation task (e.g., EU-ADR (191)). In this case, the manual annotation task is defined as the validation or curation process of annotations automatically generated. In addition, the curation process not only reviews annotations, which were automatically extracted by text mining techniques, but it, also, carefully browses the text in order to detect if they are missing (191,197).

1.5.2 Gold Standards on genome medicine and pharmacogenomics

Neves (2014) presented a comprehensive overview about corpora available in the biomedical domain, showing the limited availability of annotated corpora regarding gene and disease associations (196). Figure 1.6 shows the most popular available corpora according to the biological mentions contained, as presented by Neves (2014).

There are several corpora publicly available with annotations on genes or diseases. Table 1.1 shows a wide list of corpora focusing on the genome medicine and pharmacogenomics areas, including annotations of genes/proteins, drugs and diseases, as well as, their relationships in several cases.



Neves M 2014 [version 1; referees: 2 approved] F1000Research 2014, 3:96
(doi: 10.12688/f1000research.3216.1)

Figure 1.6. Classification of corpora according to the semantic annotations involved.

In 1999, the Genia project was launched to develop a corpus annotated with different biological entities (protein, DNA, RNA, cell line and cell type, among others) (192). The corpus, released in 2003, was developed by combining text mining and expert curation. It contained almost 100,000 annotations from 2,000 MEDLINE abstracts focused on molecular biology (198). A variety of studies based on Genia have been presented thereafter (69,89,199,200). In 2002, Franzén et al. (2002) presented the Yapex corpus as an alternative to the GENIA corpus that addressed, only, protein annotations (84). It consisted of 101 annotated MEDLINE abstracts annotated by domain experts.

It should be noted that, the above corpora did not contain any annotation of normalized entities. As mentioned in Section 1.4.3.3, the BioCreAtIvE (17) challenge initiated the identification of gene/protein names in text, by focusing on three specific species: yeast, fly and mouse (87). An annotated corpus was provided for the evaluation of text mining systems. Later, in 2007, the BioCreAtIvE II (18) was organized with a Gene Normalization (GN) task aimed at human genes (88) and an annotated corpus for human gene identification was provided (201).

There are also some corpora available with annotations on diseases. In 2004, the BioText Corpus (202) and PennBioIE Oncology Corpus (203) were the first developed corpora with disease annotations. The former included annotations of diseases and treatments, while the latter treated oncology and cytochrome P450 enzymes.

In 2009, the Arizona Disease Corpus (AZDC) was developed by two expert annotators, following specific annotation guidelines (204), in order to perform proper annotations (114). The AZDC corpus consists of 2,783 sentences from 793 MEDLINE abstracts with more than 3,000 annotations of disease names, including disease name boundaries linked to their corresponding UMLS concept identifiers. Naturally, the AZDC corpus became an important Gold Standard for development and evaluation of both, DM and DN systems (116,117).

Gurulingappa et al. (2010) presented a work focused on the identification of diseases and adverse effects in biomedical literature, and developed a new corpus called SCAI Disease corpus (103). The corpus was based on 400 randomly selected MEDLINE abstracts from a set of documents focused on diseases and adverse effects. Various experts in the field annotated 1,428 diseases and 813 adverse effects mentions without performing normalization.

Doğan & Lu (2012) described their efforts in improving the AZDC corpus by building a richer, broader and more complete diseases name corpus (39). More recently, they presented an updated corpus (NCBI Disease corpus) (122) to include concept annotations by MeSH identifiers. It was annotated by twelve experts (two experts per abstract), thereby, guaranteeing a more representative view of the disease entity. In addition, the disease mentions were classified according to the following categories: (i) *Disease Class*, (ii) *Specific Diseases*, (iii) *Composite Mention* and (iv) *Modifier*. The NCBI Disease corpus consisted of 793 MEDLINE abstracts with a total of 6,900 disease mentions identified with MeSH concepts.

With respect to the annotated corpora on gene-disease associations, it should be noted that, only a small number of works have been reported (196). Craven et al. (1999) presented a corpus with PPIs and gene-disease association annotations (mentions are not identified) (135), which was, afterwards, used for the development of systems that detect binary relationships.

Van Mulligen et al. (2012) developed a corpus with annotations between genes, drugs and diseases (linked to their corresponding identifier), by including binary associations between them (191). This corpus was composed of 300 abstracts.

Table 1.1. List of corpora involving gene/protein, drug and disease entities.

Corpora	*Year*	*Entity Mention*	*Entity Normalization*	*Relationships*	*Number of Documents*
Craven	1999	Proteins, genes and diseases	-	PPIs and gene-disease associations	1,677 abstracts
EDGAR	2000	Genes, drugs and cells	-	Gene-drug, gene-cell and drug-cell associations	103 abstracts
Yapex	2002	Proteins	-	-	101 abstracts
Genia	2003	Proteins among others	-	-	2,000 abstracts
BioText	2004	Diseases	-	-	3,655 sentences
PennBioIE	2004	Diseases	-	-	2,514 abstracts
AIMed	2005	Proteins	-	PPIs	200 abstracts
BioInfer	2007	Proteins	-	PPIs	1,100 sentences
HPRD50	2007	Proteins	-	PPIs	50 abstracts
BioCreAtIvE II GM	2007	Genes	-	-	15,000 sentences
BioCreAtIvE II GN	2007	Genes	NCBI Gene	-	543 abstracts
EU-ADR	2008	Genes, drugs and diseases	-	Drug-target, drug-disease and target-disease	300 abstracts
OSIRIS	2008	Genes and sequence variants	NCBI Gene, dbSNP	-	105 abstracts
AZDC	2008	Diseases	UMLS	-	2,783 sentences
SCAI Disease	2011	Diseases	-	-	400 abstracts
NCBI Disease	2012	Diseases	UMLS	-	793 abstracts

2 Objectives

"A goal without a plan
is just a wish"
Antoine de Saint-Exupery (1900-1944)

The application of high throughput technologies in life sciences based on the success of the Human Genome Project has led to a massive growth of biomedical data. This growth has been accompanied by a parallel increase of biomedical publications. Text mining approaches have emerged as indispensable tools to support scientists in identifying, extracting and structuring relevant biomedical data. This is particularly evident in areas such as genome medicine and pharmacogenomics, where thousands of articles are published each year.

The general aim of this thesis is to develop a text mining system for the extraction of information relevant to genome medicine and pharmacogenomics. In particular, the specific goals are:

(1) To develop a NER to identify genes and diseases from the scientific literature.

(2) To develop a RE to identify associations between genes, diseases and drugs.

(3) To apply the text mining system to different projects in the area of genome medicine and pharmacogenetics, including the development of knowledge resources.

(4) To evaluate the text mining approach in a text mining community challenge.

3 The *Befree* System

"Everyone should know
how to program a computer
because it teaches you how to think"
Steve Jobs (1955-2011)

As an objective of this thesis, the **B**io-**E**ntity **F**inder and **RE**lation **E**xtraction (BeFree) system was developed. BeFree is a text mining tool that involves two main applications related to the IE task: (i) named entity recognition (NER) and (ii) relation extraction (RE). Different techniques and approaches have been developed for each of the above applications, in order to achieve the fundamental and continuously evolving objectives of this thesis. In this chapter, it will be described how the BeFree tool addresses the above two applications (NER and RE).

3.1 The Named Entity Recognition

BeFree implements a NER based on dictionary and rule-based approaches. A dictionary-based approach focuses on the detection of biomedical entity names and a rule-based approach on the elimination of erroneous detections. In the following sections, the NER approach is explained in more detail. In particular, the biomedical text used for mining, the development of the dictionaries (for gene and disease names) and the procedure followed for the NER task are described.

3.1.1 Implementation of NER

The NER approach is mainly focused on human genes and diseases. In addition, alternative NERs can be used to detect different kinds of entities and to complement the information extracted by BeFree. For example, SETH (205) and tmChem (206) NERs have been used together with BeFree for the detection of mutation (e.g., “Gly82Ser”, “S455N”, “6310C>T” and “rs3750805”) and chemical (e.g., “Nafcillin”, “3-O-(2'-acetoxy)benzoyl-2-glucopyranose” and “aspirin”) names, respectively.

The NER is implemented in *Python 2.7.0 release* (207). Python is a high-level scripting language, thus strong in text processing. Nowadays, Python is used for a great variety of applications, in particular, for text processing. Fast tools for data analysis in Python often use C or FORTRAN in order to optimize common operations. Finally, the main characteristic of a Python code is that it is user-friendly.

Many tools and packages have been developed to assist text processing in Python. The *Natural Language Tool Kit (NLTK)* (208) is, probably, the most frequently used package in Python for human language data processing. Specific functions included in NLTK package are used for text processing in the NER.

Furthermore, the “regex” package (209) was used to create the patterns for the matching task. “Regex” is an alternative package to the Python’s current regular expression (“re”) module implementation, which is possibly the best performing regular expression engine available in a mainstream programming language. The “regex” package combines some of the advanced features of *.NET* (such as capture collections and right-to-left matching) with some of the advanced features of *Perl*, *PCRE* and *Ruby* (subroutines and recursion). The most important characteristic for the development of the NER is that ‘*regex’* allows and efficiently manages a huge list of terms, compiled in a unique and large pattern, for matching task in text.

3.1.2 Abstracts as standard input

As discussed in Introduction (see Chapter 1), scientific literature is much more accessible than other types of biomedical text (e.g., clinical reports, etc.). Another important aspect to take into account about the use of abstracts is that they describe in a concise manner the main findings of a study. This aspect allows for large scale data processing covering different studies.

Abstracts, specifically from MEDLINE, are the most commonly used type of documents for biomedical text mining applications. MEDLINE allows open access to them. In contrast, the access to the full-text article requires a license, together with a possible requirement for the processing of PDF files with its inherent difficulties. Nevertheless, there are currently several free repositories of full-text biomedical articles, such as PubMed Central (PMC) (210), where 3.8 million articles are archived. This is significantly lower number compared to those available on MEDLINE (more than 26 million abstracts). For this reason, most of the BeFree applications developed in this thesis work are based on MEDLINE abstracts.

It should be noted that, BeFree does not implement a document retrieval system (see Section 1.3.2). Therefore, the PubMed search engine is queried with the purpose of retrieving relevant publications from MEDLINE. By using a set of keywords based on MeSH terms, PubMed returns a collection of publications dealing with a specific topic(s) of interest, which can be easily downloaded.

In order to efficiently access all required abstracts, a local repository of MEDLINE has been developed. The abstracts were stored by keeping a similar XML structure of MEDLINE abstracts.

3.1.3 Dictionaries: building the basis of the NER

3.1.3.1 Defining dictionaries

BeFree implements a dictionary-based approach to detect gene and disease names from the literature. Dictionaries are large collections of terms representing biomedical entities. Additionally, each term can be linked to a unique identifier, which represents a biomedical concept.

Matching processes are applied to look-up matches between text and terms included in dictionaries and as discussed in Section 1.4.3, dictionary-based approaches are limited to finding only terms that are included in the dictionary.

Subsequently, a dictionary (e.g., dictionary of gene or disease names) should cover the maximum number of terms, which can be associated with vocabulary, as well as, with linguistic variations (e.g., abbreviations, spelling and morphological variations). In this context, dictionaries involving gene and disease entities should consist of thousands of identifiers linked to hundreds of thousands terms.

A dictionary should have a wide variability of terms representing a concept. However, different concepts could share terminology causing an ambiguity. The *ambiguity index* can be used as defined in (211) to quantify the fraction of terms that refer to different concepts and therefore identifiers, while the *variability index* quantifies the average number of terms for each concept (211). The *variability index* and *ambiguity index* of a dictionary are potential indicatives of its performance.

The dictionary-based approach implemented in BeFree covers the maximum number of terminology based on two phases: (i) by collecting the maximum number of terms for each biomedical entity (genes and diseases) from different biomedical resources that contain terminological information and (ii) by ensuring the detection of linguistic variations in the matching task.

Specifically, the matching task applied in BeFree is based on a preprocessing step of the dictionaries in order to populate them with a large extension of linguistic variations. For example, the term “breast tumor” could appear in a text as “breast tumors”, “breast tumour” and “breast tumours” or “tumor of breast” including its variations.

BeFree preprocesses all terms collected from different terminological resources in order to satisfy a wide coverage of linguistic variations. Then, the “regex” package is applied with a large regular expression, previously compiled allowing a fuzzy- or soft-matching.

In the following sections, the development and extension processes of the dictionaries of genes and diseases are described in more detail.

3.1.3.2 The gene dictionary

The gene dictionary should cover the large number of genes and proteins present in current catalogs. Three important biological databases referring to genes and proteins are used to collect the highest number of terms: (i) NCBI Gene (16), (ii) UniProt (62) and (iii) HGNC (212).

Specifically, NCBI Gene integrates gene information from a wide range of species (note that only the human species is taken in account). UniProt is a comprehensive, high-quality and freely accessible database of protein sequence and functional information. It represents a rich resource on nomenclature about proteins. HGNC is responsible for approving unique symbols and names for genes to allow unambiguous scientific communication.

These databases share many terms to define genes, where each term includes specific vocabulary and synonyms. Consequently, it is required to integrate terminology for all genes in each database.

Initially, for each database, the files containing, among other things, information about nomenclature for genes are downloaded. Files include approved or official nomenclature. An example of the files downloaded for each of the above three databases, together with the fields used is shown in Table 3.1.

Table 3.1. A list of the fields used for the extraction of gene terminology from the files downloaded for each database (NCBI Gene, Uniprot and HGNC). The fields used, contain information of vocabulary and synonyms for genes.

Database	*Filename*	*Fields*
NCBI Gene	gene_info.gz (213)	'symbol' 'synonyms' 'description' 'symbol from nomenclature authority' 'other designations'
Uniprot	uniprot_sprot_human.dat.gz (214)	'recommended name' 'alternative name' 'gene name'
HGNC	hgnc_complete_set.txt (215)	'approved symbol' 'approved name' 'previous symbols' 'previous names' 'synonyms' 'name synonyms'

By cross-reference information between databases, the terminology provided for each database can be integrated for the same gene. In particular, NCBI Gene contains cross-references to HGNC, UniProt to NCBI Gene and HGNC to the other two databases.

In more detail, a unique identifier is created to map identifiers from each biological database. This new identifier is called cross-reference unique identifier (X-RUI). An X-RUI can refer to one or multiple NCBI Gene, UniProt or/and HGNC identifiers.

Figure 3.1 illustrates an example of the integration task with reference to the Lipocalin-2 gene. Note that the same names can be used to refer to the protein (e.g., "Neutrophil-gelatinase associated lipocalin" in UniProt) and to the gene (e.g., "Neutrophil-gelatinase associated lipocalin" in NCBI Gene), since both types of entities share, in general, their terminology. Thus, for the sake of simplicity, we refer to genes and proteins as genes.

Finally, the terms and synonyms are combined and linked with the corresponding X-RUI identifier, thereby, generating a raw dictionary of genes.

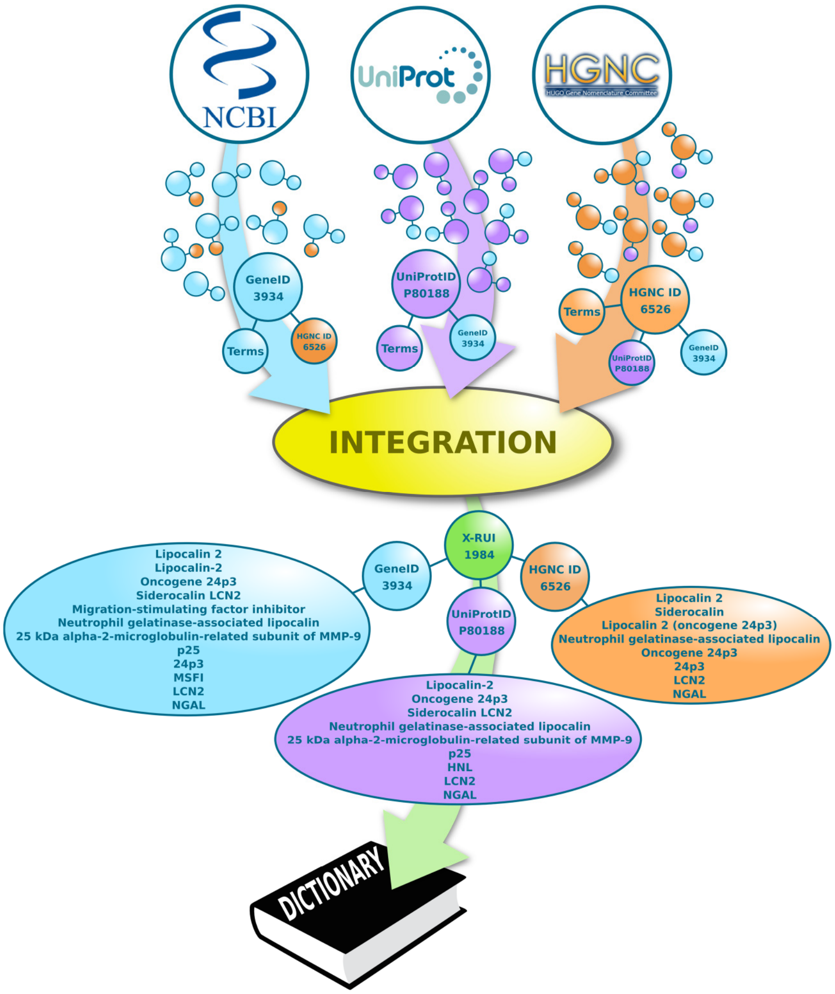


Figure 3.1. Diagram illustrating the workflow of the integration process for the extraction of gene terminology in the case of Lipocalin-2. Three databases are used (NCBI Gene, Uniprot, HGNC) and the extracted terminology is integrated into a unique identifier (X-RUI "1984").

3.1.3.3 The disease dictionary

The Unified Medical Language System (UMLS) Metathesaurus (28) is a large, multipurpose and multilingual thesaurus that contains millions of biomedical and health-related concepts, their synonymous names, and their known relationships. In particular, the UMLS covers a wide range of vocabularies related to diseases from many resources (such as MeSH, OMIM, SNOMED-CT, ICD9-CM, etc.).

For this reason, the UMLS Metathesaurus has been used as an integrated resource of disease terminology. The UMLS provides many files, which have been designed for an easy integration in a MySQL database (specifically, the UMLS 2014AB release was used).

Specifically, the *MRCONSO* and the *MRSTY* tables were used for the vocabulary extraction. The *MRCONSO* table contains records with terminological information, and the semantic type for each record is included in the *MRSTY* table (such as *'Anatomical Structure'*, *'Disease or Syndrome'*, *'Drug Delivery Device'*, *'Laboratory or Test Result'*, *'Population Group'* and *'Social Behavior'*).

First, from 133 semantic types included in the *MRSTY* table, only those related to diseases were selected. In particular: *'Congenital Abnormality'*, *'Acquired Abnormality'*, *'Disease or Syndrome'*, *'Mental or Behavioral Dysfunction'*, *'Experimental Model of Disease'*, *'Sign or Symptom'*, *'Anatomical Abnormality'*, and *'Neoplastic Process'*.

Subsequently, all records referring to the previously defined semantic types (only in English language) were selected from the *MRCONSO* table, thus, generating a raw dictionary of disease names, where terms are linked to a Concept Unique Identifier (CUI).

3.1.4 Terminology preprocessing for the matching task

From the terminology included in the raw dictionaries, erroneous or undesired terms (or characters) can be found. Furthermore, raw dictionaries do not cover all vocabulary variations used in publications, therefore presenting a major obstacle to effectively detecting biomedical entities in text.

In order to face this limitation several processes to clean (or curate), extend (or generate) and simplify (or normalize) the terminology contained in dictionaries are applied to facilitate the matching task.

In brief, a curation process “cleans” the terminology from undesirable and spurious terms. Terms normalization allows mapping variants of a term to a single, standardized form. Furthermore, terms can be processed with the purpose of automatically generating additional variants to cover a wide range of vocabulary variations that are not originally contained in the dictionaries.

Moreover, an additional process can be required before or after preprocessing methods, depending on terminological resources or the type of biomedical entity.

In this scenario, BeFree applies preprocessing methods, based on multiple rules, to both dictionaries in order to face the coverage and matching limitations.

However, before applying any defined rule, additional processes are applied in the case of the disease dictionary. As noted earlier, BeFree is focused in the detection of human genes and diseases. In this context, the gene dictionary was collected from databases containing specific gene information for humans. However, terms included for the disease dictionary were only selected by semantic types from a huge metathesaurus (not by human specification).

The semantic types do not indicate if a disease is for human or any other species, since a disease can be, typically, suffered by different species. As a consequence, specific diseases for non-human species are also included in the dictionary. Table 3.2 shows examples of disease concepts (CUIs) referring to non-human diseases.

In order to remove these diseases, an exhaustive search in the diseases dictionary was applied to detect non-human species in the terminology, to obtain a list of disease candidates to be excluded. Then, a manual inspection of the list was required to ensure an appropriate suppression of diseases. As a result, 1,939 disease concepts were defined as non-human diseases, and subsequently removed.

Table 3.2. Examples of UMLS concepts for non-human species.

UMLS Concept (CUI)	*Term*
C0276479	Mouse hepatitis
C0276477	Sialodacryoadenitis of rat
C0041307	Tuberculosis in cattle
C0275614	Ovine bihead
C2349765	Bovine stomatitis
C0271942	Goat milk anemia
C0263462	Feline acne
C0334680	Veterinary tumor
C0392661	Dog tapeworm infection

Furthermore, the *Casper* tool (216) was applied in the dictionary of diseases. Casper is a UMLS-oriented rule-based tool (for this reason, Casper was not used for the genes dictionary.). Specifically, Casper suppresses undesired terms (e.g., repeated terms) and generates additional synonyms and spelling variations (e.g., "renal hyperchloremic acidosis", "immune deficiency syndrome acquired", "parotid neoplasm", "peripheral nervous system Disorder" and "RTK" are variations generated from the following original terms: "acidosis, renal hyperchloremic", "acquired immune deficiency syndrome", "neoplasm of Parotid", "Disorder of the peripheral nervous system" and "Rhabdoid Tumor of Kidney (RTK)", respectively).

Subsequently, the disease dictionary can be preprocessed by BeFree. Table 3.3 shows the multiple rules defined in BeFree to carry out the curation, extension, generation and normalization processes. Each dictionary implements a different order of rules, according to the best *ambiguity index* and *variability index* values, similarly to the study presented by Tsuruoka et al. (2008) (211).

The terminology extracted directly from biological databases can be regarded as raw vocabulary. Some terms can be considered as undesirable (e.g., terms referring to general concepts as “receptor”, “gene”, “disease”, “infection” and “complications”), spurious (e.g., terms with less than three characters or including only numeric/punctuation characters), or they can contain some mistakes or undesirable characters (e.g., terms with annotations not used in free text as “[X]Dementia in Alzheimer's disease (disorder)”). Then, curation rules are applied ‘to clean’ the raw terminology, thereby, facilitating the matching task in the literature.

Rules for extending and generating new variations are also applied. For example, from the “IL 2r” term, replacing Arabic with Roman numbers can generate a new variant of a term (e.g., “IL II”), while an extension rule can replace “r” by “receptor” (e.g., “IL 2 receptor”).

An important step in the preparation of the vocabulary is the normalization process and is performed with the aim of simplification, such that variants of a term can be mapped to a single one (e.g., “IL 2”, “IL(2)” and “IL-2” terms are simplified to “IL 2” by removing punctuation marks).

In order to illustrate the effect of the normalization process, Table 3.4 shows the number of concepts, the number of terms, the *ambiguity* *index* and the *variability index*, before and after the normalization process for both disease and gene dictionaries.

Table 3.3. List of rules implemented in the preprocessing of terminology.

Rule	***Example***
Removing additional annotations	"[X]Gastric neurosis[Disease/Finding]" → "Gastric neurosis"
Removing incorrect terms	"23.44" or ";" → removed
Removing punctuation marks	"Interleukin-2" → "Interleukin 2"
Removing terms smaller than 3 characters	"AD" → removed
Converting to lower case	"FALDH deficiency" → "faldh deficiency"
Applying char-digit split	"chromobox5" → "chromobox 5"
Converting to Greek letters	"HP1-beta" → "HP1-β"
Converting to Roman numbers	"Interleukin 2" → "interleukin II"
Simplifying Latin letters	"Sjögren-Larsson syndrome" → "Sjogren-Larsson syndrome"
Spelling variation	"Breast cancer tumor" → "Breast cancer tumour"
Extending terms	"IL 2r" → "IL 2 receptor"

It is important to emphasize, herein, that each dictionary has its own distinctive features; for example, the gene dictionary has a high prevalence of acronyms (including numbers, punctuation marks, Greek letters, etc.) referring to genes (e.g., "A2MP1", "NOTCH1", and "SF3B1"), whereas long terms and acronyms with characters are prevalent in the disease dictionary (i.e., "Alzheimer's disease", "Acute lymphoblastic leukemia", "Primary eosinophilic endomyocardial restrictive cardiomyopathy" and "Rheumatic tricuspid stenosis and insufficiency”). Consequently, not all rules obtain the same effect in each dictionary.

The best normalization process is the one that improves the *variability index* by minimizing the *ambiguity index* of dictionaries. In the case of the gene dictionary, the number of terms between raw and curated dictionaries increases by 19% with a slight effect in ambiguity. In the case of the disease dictionary, there are no major changes in *ambiguity* and *variability indexes* after the dictionary curation process. Moreover, the final number of terms decreases significantly and this could be explained by the high number of repeated terms included in the UMLS and those generated by Casper.

Table 3.4. Showing statistics of gene and diseases dictionaries. For each dictionary, the table contains the characterization of the raw and curated version. Specifically, for the disease dictionary, the contents and statistics of the resulting dictionary using Casper are included.

Dictionary	***Number of concepts***	***Number of terms***	***Ambiguity index***	***Variability index***
Gene raw	51,429	576,784	1.53	6.12
Gene curated	51,429	705,525	1.48	13.72
Disease raw*	87,910	355,147	1.01	4.04
Disease Casper	77,366	306,009	1.01	3.96
Disease Casper & curated	77,366	298,879	1.02	3.86

*In this dictionary, the general and non-human concepts have been removed.

3.1.5 The NER process

3.1.5.1 Basic principles

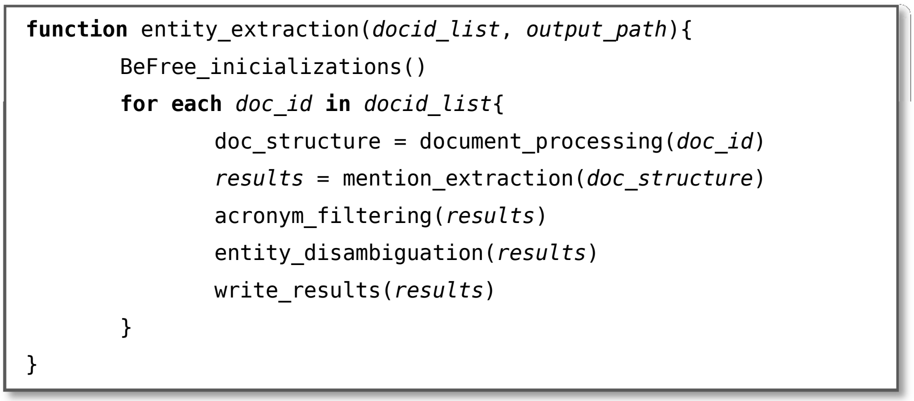
All implemented functions for the NER have been included in a Python's project called `BeFreeNER`. This project can be imported from other Python's modules using the import command.

The main function in the `BeFreeNER` project is `entity_extraction` where a list of PMIDs, an output file and the type of entity for identification, are the input parameters. Figure 3.2 shows an example of how to apply the BeFree NER for the detection of genes and diseases in a set of four abstracts.

```
import BeFreeNER
pmid_list = [‘25162549’, ‘25416513’, ‘2548710’, ‘25791637’]
output_path = “result.txt”
entity_extraction(pmid_list, output_path)
```

Figure 3.2. Example of the BeFree NER code using Python. `BeFreeNER` is , initially, imported. Next, a list of PMIDs is defined for processing. Then, a path for the output file is defined. Finally, the `entity_extraction` performs the recognition process of gene and disease names in abstracts.

The `entity_extraction` function is composed of four main steps, which process and extract information from a given abstract. Specifically, these steps are *Document Processing*, *Mention Extraction*, *Acronym Filtering* and *Entity Disambiguation* (see Figure 3.3 for a schematic pseudocode of the `entity_extraction` function). Each step is, briefly, described below. A more detailed description will be provided in the following sections.



```
function entity_extraction(docid_list, output_path){
        BeFree_inicializations()
        for each doc_id in docid_list{
                doc_structure = document_processing(doc_id)
                results = mention_extraction(doc_structure)
                acronym_filtering(results)
                entity_disambiguation(results)
                write_results(results)
        }
}
```

Figure 3.3. Pseudocode representing the `entity_extraction` function.

In the beginning, BeFree should be initialized, that is, structures and classes requiring initializations are loaded. The Sentence Splitter tool implements an algorithm that allows breaking a text into sentences. Particularly, the Sentence Splitter tool used is included in the NLTK. This Sentence Splitter implements a supervised algorithm, which needs trained data to split sentences. Specifically, the *'tokenizers/punkt/english.pickle'* model included in the NLTK package is used for Sentence Splitter initialization.

At the same time, the BeFree NER structure is initialized. BeFree collects all terminology contained in previously processed dictionaries (see Section 3.1.4). Next, the "regex" package compiles a large pattern, which will be applied in each sentence.

The Abstract Processing step retrieves the abstract information from a PMID. This information is processed and structured in order to facilitate the subsequent steps.

The Mention Extraction step implements a sequence of subroutines (see Section 3.1.5.3), which are sequentially executed on each document. This step extracts the maximum number of mentions potentially referring to gene and disease names. Furthermore, additional filtering steps are applied for the removal of erroneous (*false positive*) results, specifically, the Acronym Filtering and Entity Disambiguation steps.

Table 3.5. Description of the resulting columns of the generated text file by the recognition process.

Field	***Example***
Document identifier	PMID:26775353
Publication year	2016
Journal name	Eur. J. Gynaecol. Oncol
ISSN number	0392-2936
Abstract section	CONCLUSION
Section number	4
Sentence number	7
Mention identifier	XRUI:5587|XRUI:5310
Mention text	PKD1
Mention offset	27:31
Long term referenced before	Protein kinase D1 (in sentence #1, offset 40:57)
Sentence	The authors confirmed that PKD1 was downregulated in invasive breast cancer.

Finally, the `entity_extraction` function generates a tabular text file by default. The resulting file is composed of 12 fields, including information of the abstract (such as PMID, year of publication and name of the journal) and the mentions found (mention contained in the sentence, the offsets and the sentence). An example of a row contained in the resulting file is shown in Table 3.5.

3.1.5.2 Document Processing

In this step, the abstract information is retrieved by the PMID identifier from the local repository. Then, in order to facilitate the matching task, each abstract is processed and organized in an abstract structure.

The abstracts are, often, presented in sections (such as introduction, methods, results and conclusions) and this information is included in the abstract structure in order to indicate the section in which the NER finds information. If the abstract is not structured in sections, the NER divides the abstract in three parts, including an introduction, a body and conclusions. For example, in Section 4.1 this information was taken into consideration to score disease biomarkers associations, assuming that the title or the last part of the abstract tends to express more concisely the final message of the publication, whereas the rest of the abstract contains background information and more hypothetical discourses as contextual information of the study.

In the end, each section is divided into sentences using the sentence splitting previously mentioned. Then, each sentence is pre-processed for cleaning. This occurs, for example, in the case where the sentences have been retrieved from XML documents, such that, they can contain XML tags and thereby, leading to problems with the position of text characters (such as "&amp;", "&mgr;" and "&gt;" are converted to "&", "mu" and ">", respectively). Furthermore, some expressions can be “hidden” in the text in order to prevent errors of detection, such as units of measurement (e.g., a real number followed by “mg” or “µmol/L”) or percentage values (e.g., “8%”, “34.56%” and “.34%”).

3.1.5.3 Mention Extraction

Mention extraction is the most important step in the process of detection and identification of entities. The aim of this step is to find the maximum number of mentions that refer to a specific entity (a gene or a disease). Mention Extraction processes sentences sequentially and the information extracted is stored in a structure.

The compiled pattern (using the “regex” package), initially, processes the sentence in order to obtain a set of the longest matched mention in the Pattern Matching step. Each mention includes specific information, such as the mention text itself, a character offset determining the text’s position in the sentence and the unique database identifiers.

Next, each mention is enriched with more information based on context features. This involves a Features Enrichment step, which searches for orthographic and context features associated with the mention. For example, if the term is an acronym, it is written in plural or includes a key word for genes or diseases (such as “receptor” or “dystrophy”, respectively). These features play an important role in the detection of errors and disambiguation of entities. Table 3.6 shows the list of features that a mention can include.

Table 3.6. List of features used for the Mention Extraction step.

Feature	*Description*
SYMBOL (S)	Indicates that the mention is an acronym or symbol.
	Example 1 (PMID:25439727): *Mutations in __PEX7__, __GNPAT__, and __AGPS__, all involved in the plasmalogen-biosynthesis pathway, have been described in individuals with __RCDP__.*
LONG TERM (L)	Indicates that the mention is a long term or definition.
	Example 2 (PMID:25298246): *Our previous studies also demonstrated that AQP5 was highly expressed in __epithelial ovarian cancer__ and contributed to the progress of __ovarian cancer__.*
DICTIONARY (D)	The mention is included in the dictionary.
ACRONYM EXTRACTED (E)	The mention is an acronym, which makes reference to a long term previously mentioned. It is detected in the *Acronym Learning* step. If the acronym is in the dictionary, it is also included the DICTIONARY feature.

	Example 3 (PMID: 26770982):
	__Type II diabetes mellitus__ (__T2D__) is a chronic metabolic disorder that results from defects in both insulin secretion and insulin action.
	Example 4 (PMID: 26770982):
	__Lactate dehydrogenase A__ (__LDHA__) is one of such genes.

NUMBER OR GREEK LETTER (N)	A digit is included in the detected mention.

	Example 5 (PMID: 25230976):
	Quantitative RT-PCR analysis revealed that the mRNA expression levels for the MMACHC, PTER, **EPC2**, **ATXN7**, FHIT, **KIFAP3**, **CPEB1**, **MINPP1**, **TEX264**, **FAM107A**, **UPF3A**, **CDC16**, **MCCC1**, **CPSF3**, and **ASAP2** genes, being partner genes involved in the chimeric transcripts in the initial cohort…
	Example (PMID: 25594371):
	An __α-synuclein gene__ (SNCA) polymorphism moderates the association of PTSD symptomatology with hazardous alcohol use, but not with aggression-related measures.

GENE (Ge)	The context of the mention includes a gene keyword (such as gene, protein, receptor, promoter, target and biological marker). The location of the keyword in reference to mention is also included.

	Example 6 (PMID: 23670889):
	Association of __dopamine D2 receptor__ and __leptin receptor__ genes with clinically severe obesity.
	Example 7 (PMID: 17508011):
	Therefore, we studied the association of single nucleotide polymorphism (SNP) in the __IL-6 gene (IL6)__ promoter with plasma levels of fibrinogen, CRP and hypertension.

DISEASE (Di)	*The context of the mention includes a disease keyword (such as disease, pathology, infection, disorder, syndrome and symptom). The location of the keyword in reference to mention is included.*

	Example 8 (PMID: 26591157):
	Here we present a case of 55 year old male who presented with **lower respiratory tract infection** and clinical findings of systolic murmur…
	
	Example 9 (PMID: 26468204):
	These findings further suggest that therapeutic manipulation of S6K1 could be a valid approach to mitigate **AD** pathology.

PLURAL (P)	The mention appears in plural.

	Example 10 (PMID: 26468198):
	Adolescence is characterized by drastic behavioral adaptations and comprises a particularly vulnerable period for the emergence of various **psychiatric disorders**.
	
	Example 11 (PMID: 25533828):
	MMPs and **TIMPs** play important roles in tumor angiogenesis and invasion.

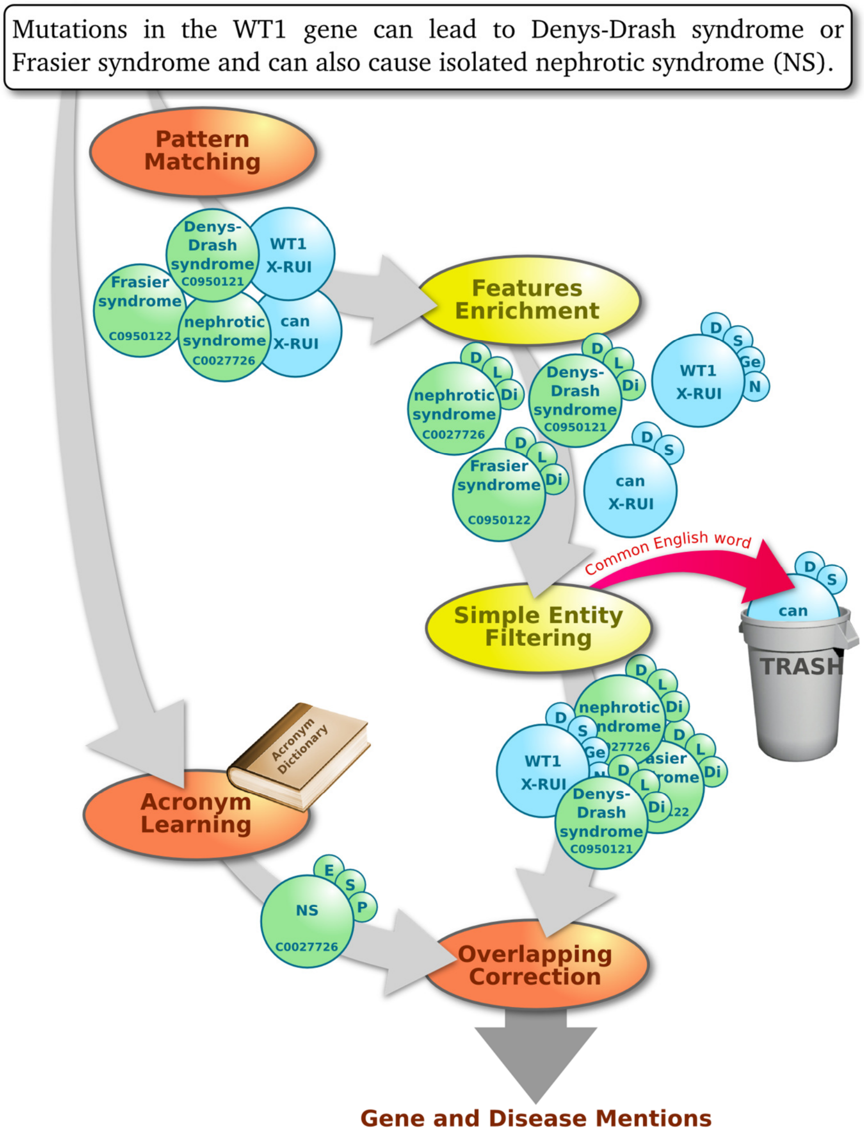
Next, a quick filter is applied to remove mentions involved in common English words (such as “can”, “back”, “are”, “but” and “full”).

Then, the Acronym Learning step helps to reference acronym terms with the long term previously mentioned in the abstract (adding the *‘ACRONYM EXTRACTED (E)’* feature, see Table 3.6). Also, by detecting an acronym-definition structure (such as *‘DEFINTION (ACRONYM)’*), acronym terms not included in the dictionaries can be detected. For example, the dictionaries do not contain terms with less than three characters; consequently they cannot be detected using the information contained in the dictionaries (see Table 3.7).

Table 3.7. Examples of the detected mentions in the Acronym Learning step, based on the "AD" and "MD" terms. In the first and second row, "AD" is detected as Alzheimer Disease and Alcohol Dependence, respectively. "MD" is detected in the rest of examples representing Major Depression, Menkes Disease and Myotonic dystrophy, respectively.

PMID	*Sentence*	*Description*
11391700	*These results are discussed in relation to neuroprotection and toxicity of the age-related pathology of **AD**.*	**AD** mention is detected as Alzheimer disease. (CUI: C0002395)
17217931	*A blunting of GH responses in abstinent **AD** men was observed only among those with the most common HTR1B promoter diplotype.*	**AD** mention is detected as Alcohol dependence. (CUI: C0001973)
16165107	*Haplotype analysis indicates that TPH-1 associates with **MD**.*	**MD** mention is detected as Major depression. (CUI: C0041696)
15923132	*The precise reasons for neurodegeneration in **MD** are poorly understood.*	**MD** mention is detected as Menkes disease. (CUI: C0022716)
10999804	*We studied the diurnal rhythmicity of cytokines and cortisol, ACTH, and dehydroepiandrosterone in 18 men with adult onset **MD** and 18 controls.*	**MD** mention is detected as Myotonic dystrophy. (CUI: C0022716)

In other words, the Acronym Learning step also detects the case where the mention is followed by an acronym or symbol, which will be always referenced to the mention in the current abstract. This acronym may not be included in the dictionary. For the purpose of detecting the new acronym in the rest of the abstract, it is included as a new acronym term in an additional acronym dictionary associated with the current abstract. This dictionary is also applied to the sentence, such that the new inferred acronyms are detected (similarly to the Pattern Matching step).



Mutations in the WT1 gene can lead to Denys-Drash syndrome or Frasier syndrome and can also cause isolated nephrotic syndrome (NS).

Figure 3.4. Workflow diagram of the Mention Extraction step implemented in the NER. The workflow begins with a sentence previously obtained from an abstract (PMID: 22763603). Pattern Matching identifies three disease mentions (in green circles) and two gene mentions (in blue circles). Each entity is enriched with orthographic and context features in the Features Enrichment step. Next, Simple Entity Filtering step removes the "can" mention. The Acronym Learning locates a new acronym mention as a disease and is included in the additional acronym dictionary and results. Finally, all mentions are reviewed by an Overlapping Correction step and the resulting mentions are returned. The processes shown in orange are executed for each sentence and those shown in yellow, are executed for each entity mention.

Finally, an Overlapping Correction step is applied in order to select a mention in the case of an overlap. In general, the longest mention is always preferred, although in some cases the selection is based on the acquired features.

3.1.5.4 Acronym Filtering

The acronyms or symbols are the most likely terms to cause errors. This is because many acronyms included in the dictionaries may represent different meanings. Accordingly, it is necessary to perform a filtering step in order to detect the incorrect acronyms found by the mention extraction step.

For example, in the sentence *'Recombinant human **OCT** expressed in E. coli was used as an antigen to obtain the monoclonal antibodies for this assay'* (from PMID: 16445902), "OCT" is correctly detected as a gene (NCBI Gene identifier "5362"). On the contrary, in the sentence *'**OCT** and red-free imaging are helpful in identifying amyloid deposits in the retina'* (from PMID: 24480837), "OCT" is not referring to a gene, but to a medical imaging technique called Optical Coherence Tomography. By applying an acronym filtering step, only the first example is reported as a correct result, filtering out the second one.

In order to carry out the Acronym Filtering step, the *Schwartz & Hearts* algorithm (217) is used in this filtering step. Schwartz & Hearts implemented a simple algorithm for identifying acronyms and their definitions focused on biomedical text. The previously described Acronym Learning step in the Mention Extraction procedure finds only acronyms for identified long terms. In comparison, the Acronym Filtering step detects all types of definition-acronym pairs (referring to biological entities or not).

The Schwartz & Hearts algorithm returns a list of acronyms-definitions pairs from a text. This list is used to detect the acronyms and definitions found by the NER that refer to other meanings.

The abstract is, initially, processed in order to obtain all acronyms or symbols with their respective long terms or definitions. If the long term, referring to an acronym, is found by Mention Extraction or is normalized by the dictionaries, the acronym is kept. Otherwise, the acronym is referred to another concept or meaning in the current abstract and thus, it is removed from results.

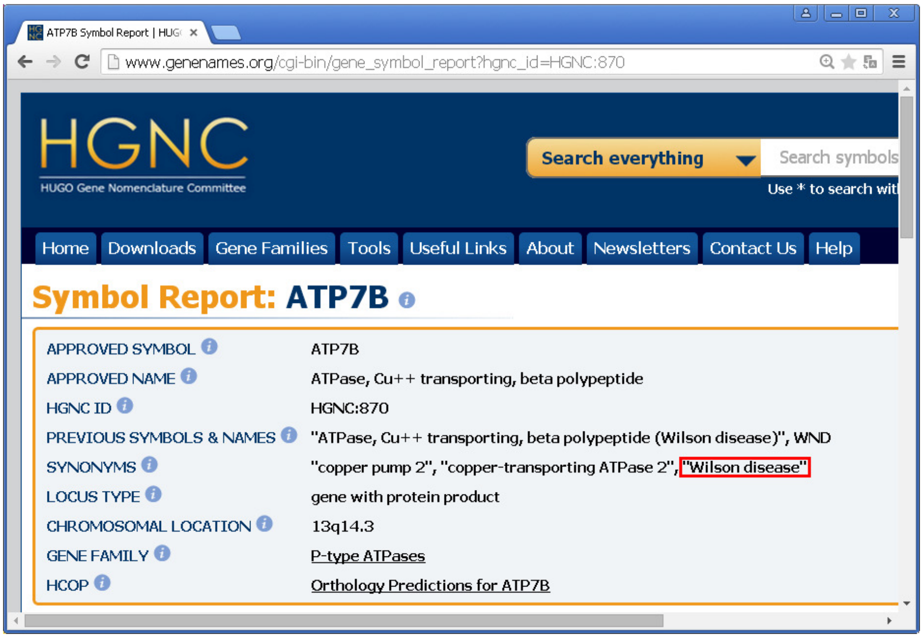
Note that the long term detected by the Schwartz & Hearts algorithm may not be exactly the same as that detected by the Mention Extraction step, but they may both share an offset overlap. Furthermore, the Schwartz & Hearts algorithm could detect a long term not included in the dictionaries that may contain additional features (see Table 3.6), such that it could be considered an entity of interest. In these cases, the entities are exhaustively reviewed, by considering the features included in the long and acronym terms during the current abstract, in order to decide which mentions refer to a biological concept.

3.1.5.5 Entity Disambiguation

Initial evaluations of the BeFree NER detected ambiguities in the normalization process and in particular, with respect to symbol names (e.g., the “NAP1” relates to at least five genes). This problem was directly addressed by using acronym definitions, which significantly improved performance. Frequently, an acronym appears after the long term is defined in the text. In this case, the list of concept identifiers of both mentions (acronym, long form) is checked using the features extracted from the *Features Enrichment* step, in order to determine if the acronym refers to the long form,. For example, in the sentence ‘Selective gene targeting using the carcinoembryonic antigen (CEA) promoter is useful in gene therapy for gastrointestinal cancer’ (from PMID: 11053994), BeFree NER detects the long form expression “carcinoembryonic antigen” as a gene with NCBI Gene Id “1048”,and the acronym “CEA” as four different gene entities (with four NCBI Gene Identifiers “1087”, “5670”, “1084” and “1048”). The concept identifier in common between the two entities (NCBI Gene Identifier “1048”) is kept as the right annotation. If there is more than one common concept identifiers, the similarity of the terms of each concept is examined in order to select the right identifier or the concept-frequency in the document.

In addition, by analyzing the initial gene-disease associations extracted by BeFree (e.g., in the work presented in the Section 4.1), an important source of ambiguity was observed, resulting from the wrong identification of entities due to ambiguities in the terminology of diseases and genes or to overlapping mentions. This is also particularly problematic in the case of acronyms, where the same symbol can be used to refer to a disease or a gene.

As commented in Section 1.4.3, there is a potential overlap between the terminology involved in both gene and disease entities. The inclusion of disease synonyms as gene names (and vice versa) in the standard databases causes an ambiguity problem between both kinds of entities. Figure 3.5 reflects this problem for the ATP7B gene in HGNC, where the "Wilson disease" term is included as a valid synonym.



Symbol Report: ATP7B

APPROVED SYMBOL	ATP7B
APPROVED NAME	ATPase, Cu++ transporting, beta polypeptide
HGNC ID	HGNC:870
PREVIOUS SYMBOLS & NAMES	"ATPase, Cu++ transporting, beta polypeptide (Wilson disease)", WND
SYNONYMS	"copper pump 2", "copper-transporting ATPase 2", "Wilson disease"
LOCUS TYPE	gene with protein product
CHROMOSOMAL LOCATION	13q14.3
GENE FAMILY	P-type ATPases
HCOP	Orthology Predictions for ATP7B

Figure 3.5. Information of the ATP7B gene provided by the HGNC database. The "Wilson disease" term (highlighted in a red box) is included as a proper synonym of the ATP7B gene.

This ambiguity problem is significant when the set of documents to be mined deals with the genetic basis of human diseases (i.e., documents including both gene and diseases names), in which ambiguity can be frequently found between genes and diseases. For this reason, the entity disambiguation between gene and disease is also addressed by BeFree, in order to identify when a mention refers to a disease or gene entity in a case of ambiguity.

The previously applied steps could, indirectly, resolve this problem in several cases. However, in this phase, it is more important to review all overlapping mentions between the gene and disease results.

Specifically, a rule-based approach is implemented to resolve the ambiguity generated between gene and disease mentions. The approach is based on orthographic and context features previously acquired by the detected mentions. If the document presents keywords representing the contained information (e.g., most MEDLINE abstracts include MeSH terms), they can be, also, used as context features.

Table 3.8 shows examples of sentences where this problem is identified. Below it will be described in more detail how the Entity Disambiguation step resolves each case.

In the first example (PMID: 23922488), two acronym mentions are found by the NER; “CHED2” is detected as a gene and a disease (NCBI Gene Identifier “83959” and UMLS CUI “C1857569”, respectively) and “SLC4A11” is only detected as a gene (NCBI Gene Identifier “83959”).

In this case, the “CHED2” mention as a gene does not contain decisive features included in the Entity Enrichment step, while the “CHED2” mention as a disease includes a fundamental feature, which refers to the "Congenital hereditary endothelial dystrophy 2" mention as the definition. This feature was obtained in the first sentence of the abstract (i.e., *'Congenital hereditary endothelial dystrophy 2 (CHED2) is an autosomal recessive disorder caused by mutations in the solute carrier family 4, sodium borate transporter, member 11 (SLC4A11) gene.'*). Consequently, the "CHED2" mention is removed from gene results.

The second and third examples show the ambiguity problem related to a long form term (i.e., "Wilson disease"). As commented before, sometimes genes are named using the involved disease name. When, the long term or definition coincides with both kinds of entities, authors tend to use some specifications. If the mention contains a *'DISEASE'* feature, as in the second example ("Wilson disease"), this is annotated as a disease. On the contrary, in the third example the mention terminates with a *'GENE'* feature, ("Wilson disease gene"), so it is annotated as a gene.

The last example demonstrates the ambiguity problem by combining, in the same sentence, the two previously described cases. The "Alzheimer disease" mention is detected as a disease and the "AD" acronym is extracted as its symbol in this abstract. However, it should be noted that, not always the "AD" mention is referred to the disease, as it happens when the mention includes, in the same abstract, a *'GENE'* feature as a last feature. In such case, the second "AD" mention found in the sentence refers to the gene involved in the Alzheimer disease.

Table 3.8. Showing four examples of sentences (last column) including ambiguity names between genes and diseases. The PMID number (first column), the detected mention (second column) and the corresponding identifiers (third column) can be seen for each sentence example.

PMID	*Mention*	*Identifiers*	*Sentence*
23922488	CHED2	GID*: 83959 CUI: C1857569	*The purpose of this study was to identify the genetic cause of **CHED2** in six Indian families and catalog all known mutations in the **SLC4A11** gene.*
	SLC4A11	GID: 83959	
6846733	Wilson disease	GID: 570 CUI: C0019202	*Serial changes of cranial computerized tomographic findings in **Wilson disease** during D-penicillamine therapy.*
15554419	Wilson disease	GID: 570 CUI: C0019202	*More than 200 mutations of **Wilson disease** gene were found, the most common ones being H1069Q (in Europe) and R778L (in Asia).*
11464541	Alzheimer disease (AD)	GID: 351 CUI: C0002395	*Almost 100 years since the first clinical report of a case of **Alzheimer disease (AD)**, three early-onset and two late-onset **AD** genes have been identified.*
	AD	GID: 351 CUI: C0002395	

*GID: Gene ID Identifier

3.2 Relation Extraction

As mentioned earlier, a key application developed on BeFree is the relation extraction (RE) between entities. Particularly, RE detects if two mentions co-occurring in the same sentence are semantically associated. In comparison with the NER, the RE implements the detection of any type of entity, such as gene-disease, drug-disease and drug-target.

3.2.1 RE based on a supervised learning approach

A co-occurrence is widely used as an indicator of association, assuming that two entities are potentially associated if they are mentioned together in the same sentence. Accordingly, for the objectives of this thesis, several co-occurrence measures have been established in order to provide a degree of association. Specifically, a score was calculated based on the part of the abstract where the association is located (title, body or conclusion) and the frequency in the literature represented (see Section 4.1).

However, the RE application implemented on BeFree goes far beyond the simple co-occurrence at the sentence-level. It incorporates a supervised learning system trained with linguistic information extracted by NLP-based methods from sentences. Specifically, the supervised learning approach processes the linguistic information involving both entities of interest (e.g., a drug and a disease) co-occurring in the same sentence, such that semantic relationships between them are detected.

The RE was developed in JAVA and uses multiple NLP-based tools methods to exploit linguistic information from sentences. On the other hand, the supervised learning approach is based on the Java tool for RE (also known as *jSRE*) (218), which is aimed at semantic relation extraction between entities at the sentence-level.

The jSRE is based on SVM, incorporating the Library for SVM (*LIBSVM*) (219,220). In addition, jSRE is released as open source (under the terms of the Apache License, Version 2.0), allowing easy modifications and new implementations.

As a rule, a supervised learning system needs corpora or annotated data "to learn" a classification task, for example, to decide if a sentence contains a real association between two candidate mentions. In order to optimize the full potential of the jSRE, multiple corpora containing sentences with annotations on entities and their relations have been processed to train models for RE.

How these linguistic features are organized, processed and learned by the supervised learning approach, in order to train a model to detect semantic relationships, is widely explained in Section 0.

In the following sections, the other two important aspects involved in the BeFree RE approach are described in more detail. Specifically, the extraction of linguistic features from sentences involving potential associations is, initially, discussed, by using NLP-based methods (Section 3.2.2). Furthermore, all models available on BeFree for detection of relationships are rigorously explained, including the corpora used for the development of each model and the type of relationship they identify (Section 3.2.3).

3.2.2 Linguistic Features

The supervised learning approach used to detect relationships between entities depends on multiples linguistic features of the sentences. These features, which are extracted by using NLP-based methods, are detailed below.

The linguistic features are based on orthographic, morphologic, syntactic and semantic features. Specifically, the orthographic features are simple annotations included in each word (e.g., if the word is formatted in lowercase, uppercase or capitalized, or if the word contains any number or punctuation mark, etc.). As semantic features, the information of target names (or role) is also considered (e.g., if this word is a gene).

Based on morphologic information, many features are also extracted and are described below:

(1) *Tokens:* are words, phrases, symbols, or other meaningful elements, as punctuation marks, that make up a sentence. A tokenizer is an algorithm or application that allows breaking a sentence into tokens, in a process called tokenization. BeFree system includes the *JULIE Lab Token Boundary Detector (JTBD)* (221) as tokenizer, a machine learning-based approach, developed and optimized for handling life scientific literature. The resulting tokens are based on words.

(2) *POS:* is a lexical category of words with grammatical properties, such as noun, verb, adjective, adverb, pronoun, preposition, conjunction and determiner. POS tagging is the process of assigning a word or token to its corresponding POS, in a sentence-context. BeFree uses the POS tagger implemented in the OpenNLP project (222) to assign a POS category to each token. This tagger is based on a supervised approach and thus, needs annotated corpus for proper operation. In this case, the PennBioIE Oncology Corpus (203) was used as training data for the POS tagger. This corpus consists of 1,414 PubMed abstracts focused on cancer, concentrating on molecular genetics and comprising approximately 327,000 words of biomedical text,

which are annotated, among other features, with the POS category.

(3) *Stem:* is the normalized form of a word that reduces all inflected forms to the same word. In other words, a stem is the part of the word that never changes even when morphologically inflected. In English, “produc-“ is the stem of the following words: “production”, “products”, “produce” and “producing”. The Porter stemming algorithm (223) is very widely known and became the standard algorithm used for English stemming, which have been included on BeFree.

(4) *Lemma:* a lemma is the canonical form or base form of a word. For instance, “be”, “are”, “is”, “was”, “were” and “being” are forms of the same lexeme, with “be” as the lemma. In comparison, “ar” and “wa” are the stems of “are” and “was2, respectively. Often the stem and the lemma can be the same for example, the words “wait” and “run2 are both lemmas and stems. BioLemmatizer is a domain-specific lemmatization tool for the morphological analysis of biomedical literature (224). This specialized tool is used by BeFree to extract lemma information from scientific text.

In addition, BeFree exploits syntactic dependencies from sentences. In particular, BeFree uses deep parsing to extract a syntactic dependency tree from sentences.

The syntactic dependency tree of a sentence consists of words linked by binary asymmetric relations called dependencies. The syntactic structure can indicate if two candidate mentions are connected (as a potential relationship) through syntactic dependencies (for more details see Section 1.4.2).

Many tools have been developed to extract syntactic dependencies from a sentence. In this case, BeFree includes the Stanford lexical parser tool (3.3.0 release), which is a well-known tool for obtaining the syntactic dependencies from a sentence as a dependency graph. This graph provides a simple description of the syntactic relationships in a sentence that can be easily processed to extract any textual relations. Figure 3.6 shows an example of dependency graph obtained by the Stanford lexical parser.

Sometimes, the syntactic dependency graph tends to be complex and provides information about many syntactic relations between words in the sentence, including candidate mentions. Using all available information provided by the Stanford lexical parser can be imprecise and has a high computing cost.

In this scenario, BeFree considers only the dependency graph involving the candidate mentions, extracting a subgraph based on *Least Common Subsumer (LCS).* In particular, this subgraph represents the shortest path between the two candidate mentions co-occurring in the sentence and the LCS is the common governor word between both (see Figure 3.6). Many works based on syntactic dependency trees have performed a similar shortest path approach (29,153,225), because long or complex sentences can be simplified in order to provide a better interpretation of the relationship between the two candidate mentions.

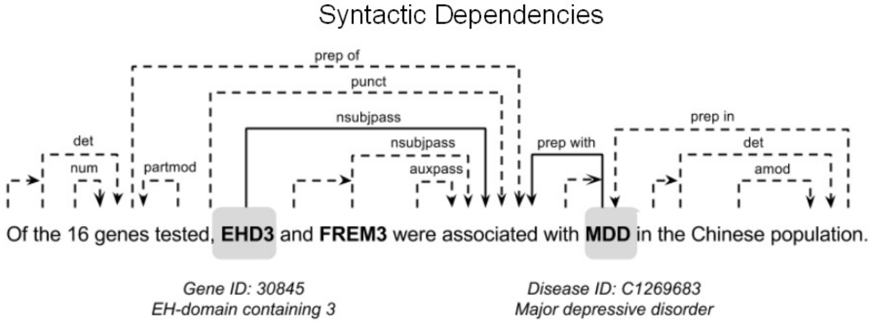


Figure 3.6.. Example of the syntactic dependency tree of a sentence ("*Of the 16 genes tested,…*"). "EHD3" and "MDD" are the candidate mentions to be related. All lines represent syntactic dependencies between words. Particularly, the solid lines represent the subgraph between the terms "EHD3" and "MDD", while the term "associated" denotes the LCS.

3.2.3 Models for detection of different associations

In comparison with the BeFree NER (which focuses on genes and diseases), the RE approach can detect multiple types of relationships. This happens because the BeFree RE depends on the training data (i.e., corpora). For this reason, BeFree RE has been trained with different corpora in order to detect multiple types of relationships. Specifically, BeFree RE is able to extract semantic relationships between genes, drugs and diseases. Most recently, BeFree RE was also used to detect chemical-induced disease relations.

Models available in BeFree are detailed below. Table 3.9 shows the list of all available models, including the type of relationship detected. In addition, the performance obtained by 10-fold cross-validation is also shown.

Table 3.9. Performance for each available model in BeFree based on the following metrics: *precision (P)*, *recall (R)* and *F-score (F).*

Corpus	***Description***	***Performance****		
		P	***R***	***F***
EU-ADR	Drug-target	74.2	97.2	83.3
	Gene-disease	75.1	97.7	84.6
	Drug-disease	70.2	93.2	79.3
GAD	Gene-disease F/T	77.8	87.2	82.2
	Gene-disease F/P/N	66.0	73.8	69.6
LHGDN	Gene-disease classification	84.7	86.1	85.4
Crow-CID	Chemical-induced Disease	82.0	73.4	76.8

*The results of 10-fold cross validation.

3.2.3.1 EU-ADR corpus

The EU-ADR corpus is a Gold Standard that contains annotations of different entities (drugs, diseases, and genes/proteins) and the relationships between them (191). In particular, the EU-ADR corpus is divided in three datasets containing annotations of relationships between drugs and diseases (drug-disease set), drugs and their protein targets (drug-target set) and genes and their association to diseases (gene-disease set). In addition, each relationship is classified according to its level of certainty, i.e.: *positive association (PA)*, *negative association (NA)*, *speculative association (SA)* and *false association (FA)*. Table 3.10 demonstrates a few examples of the levels of certainty considered in the EU-ADR corpus.

Table 3.10. Examples of the association types considered in the EU-ADR corpus.

Association Type		*Examples*
True	Positive	*'**Vascular endothelial growth factor gene** is associated with increased risk for **Alzheimer's Disease**.'* *'The **DAOA gene** has been found associated with **schizophrenia**.'*
	Speculative	*'The results presented suggest that **DDC** may act as a minor susceptibility gene for **bipolar affective disorder.'*** *'Our results suggest that **HCN1** protein could be a potential target for treatment of **anxiety** and **depression disorders**.'*
	Negative	*'These results suggest that **SLC25A12** is not a major contributor to **autism** risk in these families.'* *'Additionally, the **PKLR** and the NOS1AP genotypes were demonstrated not to have a major influence on **diabetes**.'*
False		*'We report association of **ZNF804A** with schizophrenia and CACNA1C with **bipolar disorder**.'*

3.2.3.2 GAD corpus

The Genetic Association Database (GAD) is an archive of human genetic association studies of complex diseases, including summary data extracted from publications on candidate gene and GWAS studies (226).

The information contained in GAD was used for the development of an annotated corpus composed of associations between genes and diseases (downloaded on January 21st, 2013).

Specifically, GAD contains more than 130,000 records with information about specific gene-disease associations. For the work under consideration, a record of GAD can provide the following information: (i) a gene identifier (NCBI GeneID), (ii) a disease name (not identifier), (iii) if the gene is positively (Y) or negatively (N) associated with the disease and (iv) a piece of text supporting the associations.

However, not all records provide the previously described information. Therefore, only records including information of interest were considered (it should be noted that, the piece of text should be in the form of a sentence).

An interesting point is that GAD does not provide information about boundaries of gene and disease names in the piece of text, which are a requirement to train a model. Consequently, the BeFree NER was used to identify genes and diseases in sentences.

The development of the dataset from the GAD database is described below. The disease term provided by GAD is, initially, normalized to its corresponding disease concept using dictionaries included in BeFree (with the purpose of obtaining all terms associated with the disease name). Next, for each record, the BeFree NER is applied for the detection of gene and disease mentions in the sentence. Subsequently, if both entities are identified by BeFree, the record is annotated as a true association (positive and negative). If the entities identified in the sentence are not supported by the record, then, the sentence is annotated as a false association. Otherwise, the record is discarded (e.g., only one entity was detected).

Finally, a set of 5,329 sentences was collected from the GAD database, containing 2,800 sentences with gene-disease associations (including 1,833 positive and 967 negative associations) and 2,529 sentences with false associations. The dataset was used to train a BeFree model to detect gene-disease associations.

In comparison with the EU-ADR corpus, the annotated corpus developed from GAD database contains a high number of examples reporting negative associations. Therefore, this corpus was also used to train a model to specifically detect between positive and negative gene-disease associations.

3.2.3.3 LHGDN corpus

LHGDN (Literature-derived Human Gene-Disease Network) is a text mining derived database with focus on extracting and classifying gene-disease associations with respect to several biomolecular conditions. Specifically, the textual source utilized here originates from the Entrez Gene's GeneRIF (Gene Reference Into Function) database (227).

LHGDN was created based on a GeneRIF version (March 31st, 2009), consisting of 414,241 phrases. These phrases were further restricted to the species Homo sapiens, which resulted in a total of 178,004 phrases. LHGDN provided, overall, 59,200 distinct gene-disease associations grouped in four classes related to several biomolecular conditions, such as *Genetic Variation*, *Altered Expression*, *Biomarker* and *Post-Translational Modification* (see Table 3.11).

Table 3.11. Examples of the association types considered in the LHGND corpus. The last column shows both the number of associations included in the original LHGND and the final dataset.

Association Type	***Example***		***Number of associations***	
	GeneID CUI	***Sentence***	***Database***	***Final dataset***
Genetic Variation	540 C0019202	*The researchers found a new mutation that is associated with Wilson's disease.*	18,611	15,415
Altered Expression	4221 C0025267	*MEN1-mediated caspase 8 expression in suppressing multiple endocrine neoplasia type 1 is reported*	20,890	15,965
Biomarker	6331 C1142166	*Another gene other than the SCN5A may be associated with Brugada syndrome.*	18,582	12,908
Post Translational Modification	605 C0079773	*Promoter hypermethylation of BCL7a is associated with cutaneous T-cell lymphoma*	1,117	1,034

Records of the LHGDN data source are composed of the following fields: (i) the gene identifier from NCBI Gene, (ii) the disease identifier from UMLS, (iii) the association type involved in both entities (e.g., *Genetic Variation*) and (iv) the sentence supporting the evidence. As in the case of the GAD dataset (see Section 3.2.3.2), the exact location of the gene and disease mentions in the sentence is not available, such that the BeFree NER was also used to identify them.

Table 3.11 shows the resulting number of associations (see last column) extracted by BeFree from the LHGDN database. As it can be noted, BeFree cannot extract all gene and disease mentions from all sentences included in the LHGDN database. There are several reasons that could explain this decrease:

(1) *False positives* corresponding to the BeFree NER.

(2) Identifiers included in LHGDN database can be deprecated; consequently, BeFree cannot properly normalize the respective mentions.

(3) Not all sentences from the LHGDN database that supports the evidence contain the reported entity concepts. For example, the first sentence seen in Table 3.11, contains only a mention for the UMLS concept “C0019202” (i.e., “Wilson’s disease”).

(4) A UMLS concept reported by the LHGDN database is, sometimes, used to cover a set of related disease concepts. For example, BeFree has not detected any mentions related to the “C1458155” UMLS concept (i.e., “breast neoplasm”) on 3,007 occasions. In this case, the ‘breast cancer’ mention, which corresponds to the “C0006142” and “C0678222” UMLS concepts, is identified with the “C1458155” UMLS concept by the LHGDN database. Fortunately, in order to deal with this issue, some records reporting the problem were automatically assigned with “complementary” UMLS concepts. Following the example, after correction, BeFree reported properly 3,003 sentences from 3,007 related to the “C1458155” UMLS concept.

3.2.3.4 Crowd-CID relation corpus

The crowd-CID relation corpus is a novel dataset developed in order to address the Task 3.B of BioCreAtIvE V challenge (BC5) that dealt with Chemical-induced Disease (CID) relations.

The development involved a combined approach between text mining and manual annotation. Initially, the DNorm and the tmChem NERs were used to identify disease and chemical mentions with their corresponding MeSH identifiers (which were a requirement in the task). Subsequently, sentences including at least one co-occurrence between a chemical and a disease mention were selected for the annotation task. The annotation task was based on a novel crowdsourcing-approach with the purpose of adding manual annotations to the corpus. More details about the crowd-CID relation corpus development can be found in Section 0.

The curators annotated sentences with chemical-disease associations as true examples, if the disease was caused by the chemical. Otherwise, associations were annotated as false examples (e.g.,chemical-treatments relationships).

4 Applications and results

“Persistence guarantees

that results are inevitable.”

Paramahansa Yogananda (1893-1952)

4.1 A knowledge-driven approach to extract disease-related biomarkers from the literature

Thousands of biomolecules are being investigated as potential biomarkers. The results of the research conducted are, widely, reported on the biomedical literature. In particular, genomic biomarkers together with disease-related information are, frequently, studied. Thus, the biomedical literature contains valuable knowledge for those interested in gathering information on biomarkers. In order to identify, extract, and analyze this information from literature, automatic processing of the text sources is required. With this objective, a knowledge-driven text mining approach is presented, for the extraction of disease-related biomarker information from the literature. Specifically, this approach implements a named entity recognition method to identify genes and diseases of interest and co-occurrence-based statistics methods to detect disease-related biomarkers relations. As a result, the information extracted is integrated into a new disease-related biomarkers database.

Bravo À, Cases M, Queralt-Rosinach N, Sanz F, Furlong LI. *A knowledge-driven approach to extract disease-related biomarkers from the literature.* BioMed research international. 2014 Apr 16; 2014.

Bravo À, Cases M, Queralt-Rosinach N, Sanz F, Furlong LI. A knowledge-driven approach to extract disease-related biomarkers from the literature. Biomed Res Int. 2014;2014:253128. doi: 10.1155/2014/253128

4.2 Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research

Due to the increasing size of literature repositories, there is a strong need for tools that identify and gather the relevant information from the literature and place it in the context of current biomedical knowledge. In the past, most efforts in text mining of relationships have been devoted to the identification of interactions between proteins. In contrast, less attention has been paid to the identification of relationships involving other biomedical entities. In this context, the BeFree system is presented with the aim to identify relationships between diseases, drugs and genes, with a special focus on genes and the associated human diseases. BeFree implements a NER approach, based on the previously published work by Bravo et al. (2014) (see Section 4.1) and a RE approach, based on the exploitation of semantic and morpho-syntactic information from text. BeFree is assessed based on two real-life scenarios. Finally, the impact of this approach on translational research is widely discussed.

Bravo À, Piñero J, Queralt-Rosinach N, Rautschka M, Furlong LI. *Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research.* BMC bioinformatics. 2015 Feb 21; 16(1):1.

Bravo À, Piñero J, Queralt-Rosinach N, Rautschka M, Furlong LI. Extraction of relations between genes and diseases from text and large-scale data analysis : implications for translational research. BMC Bioinformatics. 2015 Feb 21;16:55. doi: 10.1186/s12859-015-0472-9.

4.3 Combining machine learning, crowdsourcing and expert knowledge to detect chemical-induced diseases in text

This work addresses the goal of the BioCreAtIvE V community challenge (BC5), particularly, the Chemical-induced Disease (CID) relations task. This task involves the identification of diseases and chemicals to promote the development of text mining solutions for the study of drug side effects. In addition, BC5 CID task constitutes a new challenge since the relations are mentioned both at the sentence and at the whole-document level (i.e., spanning several sentences). In this scenario, a new text mining system is described for the identification of drug side effects from the literature. It consists of three approaches: BeFree RE, rule- and knowledge-based. For this purpose, a novel Crowd-CID relation corpus is developed to train BeFree. In the final evaluation setting, the system achieved the highest *recall* of the challenge (63%). By performing an error analysis, the main causes of misclassifications are identified. The need of employing consistent gold standards is highlighted, for the advancement of the state-of-the-art in text mining of drug side effects.

Bravo À, Li TS, Su AI, Good BM, Furlong LI. *Combining machine learning, crowdsourcing and expert knowledge to detect chemical-induced diseases in text.* Database. 2016 May 10; 2016:baw094.

Bravo À, Li TS, Su AI, Good BM, Furlong LI. Combining machine learning, crowdsourcing and expert knowledge to detect chemical-induced diseases in text. Database (Oxford). 2016 Jun 15;2016. pii: baw094. doi: 10.1093/database/baw094.

4.4 Text mining and expert curation to develop a database on psychiatric diseases and their genes

During the past years, there has been a growing interest in the genetics of psychiatric disorders, the findings of which have been reported on hundreds of thousands of publications. In the following work, BeFree is applied to a large set of publications, in order to extract relationships between genes and psychiatric disorders. A curation workflow is, next, implemented for the validation of the text-mined information by experts. As a result, the curated information is used to populate PsyGeNET, which is curated resource for the exploratory analysis of psychiatric diseases and their associated genes.

Gutiérrez-Sacristán A, Bravo À, et al. *Text mining and expert curation to develop a database on psychiatric diseases and genes.* Proceedings of the 7th International Symposium on Semantic Mining in Biomedicine, SMBM 2016, Potsdam, Germany, August 4-5, 2016. p.48-55.

Gutiérrez-Sacristán A, Bravo À, Portero M, Valverde O, Armario A, Blanco-Gandía MC, Farré A, Fernández-Ibarrondo L, Fonseca F, Giraldo J, Leis A, Mané A, Mayer MA, Montagud-Romero S, Nadal R, Ortiz J, Pavon FJ, Perez E, Rodríguez-Arias M, Serrano A, Torrens M, Warnault V, Sanz F, Furlong LI. Text mining and expert curation to develop a database on psychiatric diseases and their genes. Dins: Proceedings of the 7th International Symposium on Semantic Mining in Biomedicine. Potsdam, Germany, August 4-5, 2016

5 Discussion

“A person who never made a mistake
never tried anything new”
Albert Einstein (1879-1955)

5.1 Overview

The unstoppable growth of articles published in the biomedical domain has generated a large accumulation of literature, containing valuable information for the scientific community. In order to automatically extract relevant information, various text mining techniques have been applied in biomedical text sources. For this purpose, several works and tools have been presented to address specific text mining challenges (see Section 1.4), such as NER (e.g., the detection of genes or the identification of adverse effects) or RE (e.g., the extraction of gene-disease associations or chemical-induced disease relations).

In this thesis work, the BeFree system was developed as a text mining system for the extraction of relevant information from the biomedical scientific literature. BeFree faces all challenges posed by Information Extraction, both at the named entity recognition and relation extraction levels, mainly for human genes and diseases. The presented system identifies gene and disease mentions in the literature and addresses the problem of entity ambiguity. In addition, BeFree detects relationships between genes, diseases and drugs and provides information about the semantics of relationship between entities. Furthermore, BeFree can identify other types of relationships, such as the side effects produced by a drug (drug-disease relationship) and the targets of the drug (drug-gene relationship).

An essential requirement for text mining solutions in the biomedical domain is the availability of annotated corpora. However, for certain IE tasks, there are no available corpora. Therefore, several corpora were developed that addressed the emerging challenges throughout this doctoral work.

BeFree has contributed with novel structured biomedical data from the literature to populate knowledge resources. The first application was on the identification of disease-related biomarkers from MEDLINE abstracts. The results are available in a publicly available database (see Section 4.1). Second, BeFree is currently used to populate DisGeNET, a database on human diseases and their genes. BeFree contributes with more than 90% of the gene-disease associations integrated in DisGeNET (see Section 0). In addition, the data provided to DisGeNET has been also published in the PubAnnotation project (228). Third, BeFree is also used to populate the PsyGeNET database, which focuses on psychiatric disorders and their genes (see Section 0). Specifically, the information provided by BeFree was curated by a team of experts following specifically defined guidelines. Fourth, we have participated in one of the most important text mining community challenges, that is, BioCreAtIvE V (see Section 0). In particular, we have participated in the chemical-induced diseases (CID) relations task designed to advance the state-of-the-art in the identification of drug side effects from the biomedical literature.

5.2 The ambiguity problem in NER

During the initial phase of this work, several NER tools were evaluated on the basis of their ability to detect and normalize genes and diseases and to handle the ambiguities between these entities.

At that time, only Metamap (120) was able to detect and normalize gene and disease names. However, although the vocabularies used for the detection of diseases were complete and allowed detection of different types of diseases, the coverage of gene entities was quite limited.

Thus, the BeFree NER was designed based on a comprehensive dictionary of gene and disease names, developed by integrating different databases and a semi-automatic curation process was, subsequently, performed (see Section 3.1.3).

The ambiguity of entities is an issue widely described in the biomedical text mining domain (229). Different types of ambiguity can be found in a named entity (for details see Section 1.4.3). For instance, it has been reported that 85.1% of mouse genes are ambiguous with other gene names (61). Most of the works addressing the ambiguity are focused on the gene name disambiguation. Today, the correct recognition of gene names remains a challenging task due to the sharing of terminology between genes across species but also in a single species, with common English words and other biomedical terms (e.g., disease names).

In the context of the BioCreAtIvE II challenge, different approaches were used for the gene name disambiguation problem, such as the incorporation of background knowledge (90), the use of alternative dictionaries (91) and context-based approaches (92). Moreover, Word Sense Disambiguation (WSD) approaches, based on supervised learning, have been also developed, achieving competitive performance. For instance, Joshi et al. (2005) (230), Stevenson et al.

(2008) (231) and Leroy et al. (2005) (232) combined different linguistic features for training and reported a performance over 86% of *F-score* (using the same corpus annotated with unambiguous entities (233)). However, there are few corpora including a large number of unambiguous annotated entities, covering a small number of names and senses (233,234).

An important ambiguity problem regarding genes appears when the same gene name is used in different animal species. Wei et al. (2015) presented a supervised learning approach based on CRF (97). Their system exploited shallow and context linguistic information for gene name disambiguation across species. Particularly, in the evaluation of cross species gene normalization (BioCreAtIvE III GN task), this system achieved 50.1% of *F-score*, showing that gene name disambiguation across species is still an unsolved problem.

The NER module in BeFree was implemented based on dictionaries and a set of rules designed to disambiguate gene and disease names. Also, special emphasis was given to the development of a comprehensive dictionary for each type of entity, by generating linguistic variations and synonyms. The BeFree NER was evaluated on the BioCreAtIvE II GN corpus. When the disambiguation rules were not used, the NER achieved a high *recall* but a poor *precision* (*P*: 48.1% *R*: 80.1% *F*: 60.1%). The performance was considerably improved when the disambiguation rule-based approach was applied (*P*: 74.0%, *R*: 76.2%, *F*: 75.0%).

Although, the disease name recognition has received less attention than the gene name recognition in the text mining field, some works have been presented to identify disease names (see Section 1.4.3.4). One of the most recent works, DNorm (104), is based on BANNER to detect disease names and applied a supervised learning approach to compute similarities between mentions and concept names. It achieves a high *F-score* (80.9%) in the NCBI disease corpus.

In the case of detecting and identifying disease names, the BeFree NER was evaluated using the AZDC corpus and achieved competitive results (*P*: 72.1% *R*: 64.4% *F*: 68.0%) compared to previous approaches (114,116).

One of the main areas of application of BeFree in this thesis was the extraction of relationships between genes and diseases. The analysis of the first results of gene-disease associations extracted by BeFree (e.g., in the work presented in Section 4.1) showed that an important fraction of errors in RE were due to the ambiguity of entities. For this reason, it was very important to address the ambiguity between genes and diseases at the NER level to improve the results of the RE step. To the best of our knowledge, this ambiguity has not been addressed by any other text mining tool. In order to address this problem, the BeFree NER was extended with a set of rules to disambiguate between genes and diseases. Although a formal evaluation of the entity disambiguation step was not performed due to the lack of a suitable corpus, non-formal evaluations by the PsyGeNET and DisGeNET teams indicated that the results were considerably improved after the implementation of the disambiguation step.

5.3 The desirable corpus for text mining applications

Annotated corpus is an important requirement to develop and evaluate text mining systems. In recent years, many successful efforts have been made to develop and contribute with novel corpora in the biomedical domain. However, the majority of them involves only gene/protein annotations (see Section 1.5) and only two corpora on gene-disease associations were available at the beginning of this thesis work (135,191).

In order to overcome this limitation, four different annotated corpora were developed in this thesis, by using different strategies, namely: a) automatic annotation of data extracted from a database curated by experts; b) automatic annotation of data extracted from the literature (not curated by experts); c) curation of text-mined data by a group of domain experts; d) curation of text-mined data by a large group of citizens. Table 5.1 shows a comparison of the corpora contributed in this thesis.

Table 5.1. Comparison of corpora developed during this thesis. Note that EU-ADR was not developed in this work; however, it is included for comparison with a gold standard.

Corpus	*Initial document set*	*Associations*	*Curators*	*Work by Curator*	*Method*	*Extra Cost*
EU-ADR	300 abst.	941	5	300 abts.	Expert	-
GAD	~130K records	5,329	-	-	Automatic	-
LHGDN	59,2K records	45,322	-	-	Automatic	-
Crowd-CID relation	3K abst.	3,068	134	~150 sent.	Crowd workers	$763.92
PsyGeNET	~1M abst.	2,507	20	~400	Expert	-

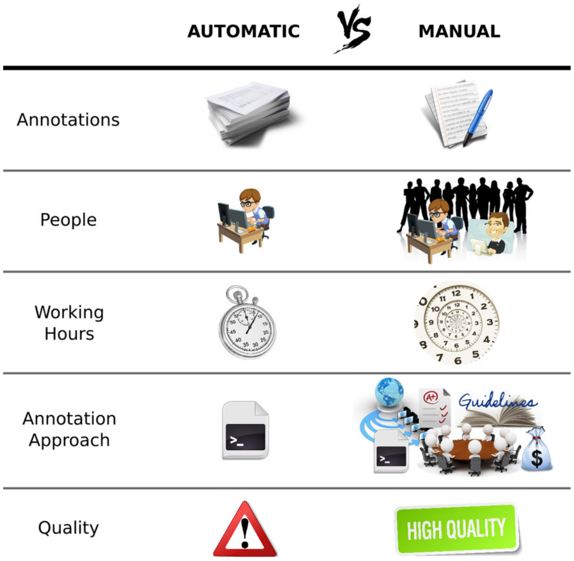


Figure 5.1. Comparison between the development of manual and automatic corpora.

The development of a corpus curated by domain experts requires a larger effort and more resources than a corpus annotated by text mining (see Figure 5.1). An increasingly adopted approach is to combine automatic annotation with expert curation (168,235,236). In this manner, the expert curation workload is partially reduced, by providing annotations obtained by text mining systems. The curators can then focus on curation tasks that are more difficult to achieve accurately in an automatic manner. The use of user-friendly curation tools is also an important asset to aid the curators' work. The development of text mining corpora by non-expert curators or "citizen scientists" has been proposed also as an alternative approach (see Figure 5.2) (237).

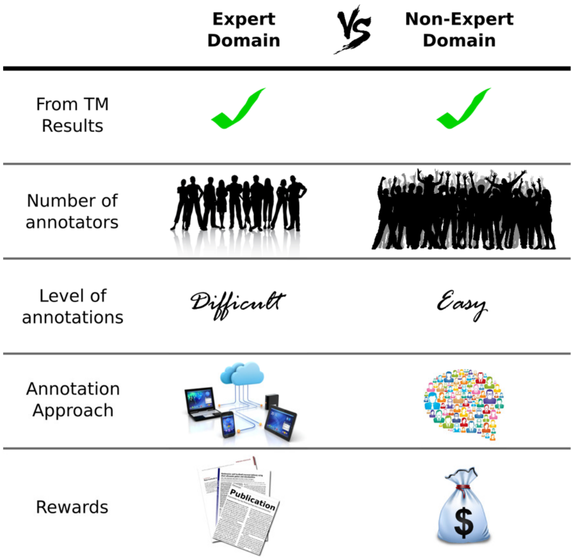


Figure 5.2. Comparing annotation processes between domain-expert curators vs. non-expert curators.

The above mentioned corpora was used to train the RE models for the identification of different types of relationships in the context if this thesis. As discussed in Section 0 and 4.3, different quality of annotations can lead to differences in performance in the RE step. For example, in the case of gene-disease associations, a model trained on the EU-ADR corpus (small in size but annotated by domain experts) exhibited a superior performance than a model trained on GAD (large in size but annotated by text mining from a database). However, it should be noted that, the model trained on GAD could still achieve a competitive performance. In addition, due to its larger size and balance between negative and positive examples, it allowed the development of a system capable of detecting both positive and negative associations.

5.4 BeFree for accurate identification of a variety of biomedical relationships

In this thesis, a RE approach was presented based on kernel methods that exploit both shallow and deep syntactic information. The results described in Chapter 4 demonstrated that a kernel-based approach, leveraging both shallow and deep syntactic information, performs competitively for the identification of drug-disease, drug-target and gene-disease relationships from free text. However, several differences were observed in the performance depending on the association type to be identified. For example, while for the identification of drug-disease and gene-disease associations the best performance was achieved (in terms of *F-score*) by the K_{DEP} kernel alone, for drug-target associations the best performance was obtained by combining both kernels. In the case of gene-disease associations, where the GAD corpus was used instead of the EU-ADR corpus, the best results were achieved with the K_{SL} kernel, contrasting with the results obtained by using the EU-ADR corpus.

It should be noted that, a RE system trained on shallow linguistic information alone is enough to produce competitive results on the associations considered in this thesis. This is an important result when considering the type of approach to be selected for a relation extraction task, in particular for large-scale analysis and/or when limited computing power is available.

As reported in Section 0 (Section 0, Table 1), the performance of BeFree on the three relation types was on par of the recent work in the field. It is worth noting that the studies cited in Table 1 defined the relationships in different ways and used different benchmarks and sometimes different metrics for evaluation. Therefore, the results of these comparisons need to be taken with caution. Unfortunately, there are no community challenges in this area (with the exception of the Drug-Drug interaction challenge (197) and the BioCreAtIvE V CID task, see below) to systematically and comparatively evaluate each of the RE tasks addressed in this thesis.

To partially circumvent this limitation, the performance of another RE system was evaluated on the EU-ADR corpus and compared with the the performance of BeFree. More specifically, the SemRep system (142,238) was adapted for the identification of the three relationship types as defined in our RE task. SemRep identified these relationship types with high *precision* but lower *recall* than BeFree. Thus, compared to SemRep, BeFree achieved more balanced results in terms of P and R for the identification of the three entity types.

The curation work performed by the PsyGeNET curation team on the gene-disease associations extracted by BeFree also allowed the estimation of the *precision* of BeFree RE. Using the data curated by the experts as a gold standard, BeFree achieved a high *precision* value (82%), leveraging only on shallow linguistic information.

In addition, the performance of the RE was evaluated for the identification of protein-protein interactions and thus, it was compared against a larger body of literature. By using the AIMED corpus (239), the use of shallow linguistic information, as well as, syntactic information in the form of dependency walk features, lead to more accurate models for PPI relation extraction. The results were comparable to those obtained with state-of-the-art approaches that were tested on the AIMED corpus (163).

The drug-disease relation extraction task was also evaluated in the BioCreAtIvE V challenge. The BeFree model trained on the crowd-CID relation corpus, a corpus created by crowd workers, achieved a high performance by 10-fold cross-validation (76.82% *F-score*). However, when assessing the performance of BeFree on the corpus provided by the task organizers, it dropped to 45% *F-score*, significantly lower than the one obtained on the crowd-CID relation corpus and on the EU-ADR corpus. These evaluation exercises, clearly, showed that the performance obtained by cross-validation often differ significantly from evaluations performed using independent data sets in real case scenarios. The error analysis performed on the BeFree results allowed the identification of some of the weak points of the system, such as the incorrect handling of negations, poor performance on long and complicated sentences with several potentially related entities, and incorrect distinction between therapeutic indication of the drug and SEs. Note that due to the processing time constraints of the BC5 challenge, only shallow linguistic features were used to train the system. Thus, in future work we plan to evaluate if syntactic dependency features improves these results.

Based on the results presented in this thesis and from works presented by others (141,144,152,153,160,162,240), we conclude that dependency features can improve a supervised learning approach, in this case based on kernel methods, for a variety of relationship types. However, shallow linguistic parsing methods are more widely extended (there is a great variety of automatic tools for shallow linguistic analysis), cover a major range of natural languages and produce faster results with smaller error rates (with complex sentences, dependency parsing is less accurate) (157,240). Thus, we propose that shallow parsing approaches are an adequate alternative in terms of computational cost and performance.

5.5 BeFree as a tool to identify actionable information

Throughout this thesis, in addition to evaluating the performance of the RE system based on *precision (P)*, *recall (R)* and *F-score (F)*, that is common practice in the text mining domain (see Section 1.3.5), the ability of the BeFree system was assessed to identify information that could be used to answer real biomedical research questions.

The ability of BeFree to identify information useful for biomedical research is illustrated by the following examples. First, gene-disease associations extracted by BeFree are used to populate DisGeNET (241,242), one of the most complete databases on human diseases and their genes. As of September 2016, the database received more than 35,000 web users and has been used in different research projects and data mining companies. Of note, gene-disease associations extracted by BeFree represents more than 90% of the data available in this database. Another example was the extraction of information for the PsyGeNET platform (243,244). In this case, the gene-disease associations extracted by BeFree were subjected to curation by a team of 22 experts to collate the data that was, finally, used to populate the database. It is worth noting that only 18% of the data identified by BeFree was judged as errors by the consensus of two experts. Altogether, this data indicated that BeFree was able to extract meaningful information from MEDLINE abstracts, regarding the genetic basis of human diseases.

5.6 BeFree captures different facets of the relations

In biomedicine, a relationship between two entities (e.g., between a gene and a disease) can be considered from different facets or perspectives. For example, a relationship can be unqualified or not specified at the semantic level (e.g., “The LOXL1 gene *is associated with* exfoliation glaucoma”), or, on the other hand, semantically specified (e.g., “The LOXL1 gene *is overexpressed in* exfoliation glaucoma”). It is evident that the second option will be the preferred one by a researcher, as it gives more information about how the gene relates to the disease, thereby, providing clues on the disease pathogenesis. The relationships can be, also, considered from the perspective of their level of certainty; that is, if the scientific statement is phrased as a proven experimental observation or fact or, alternatively, as a speculation or hypothesis (e.g., “The LOXL1 gene *might be associated with* exfoliation glaucoma”). Research in the area of discourse analysis has been applied to approach this latter perspective of RE (245–247). Finally, identification of negative findings from the literature, although not frequently reported, is also an important goal.

In this thesis, various perspectives to classify a particular relation between two entities were addressed. The initial aim was to distinguish between true and false associations, without taking into account neither the level of certainty nor a more granular description of the association type. Then, once a suitable corpus was developed, a RE system capable of identifying negations from true assertions was presented. In the particular case of gene-disease associations, a system to identify different association types was developed, according to the DisGeNET association type ontology. Finally, the identification of negative findings from the literature was addressed for the PsyGeNET project. Interestingly, 30% of the gene-disease associations, validated by experts, had at least one publication reporting negative findings on the association between the gene and the disease. This highlights the importance of identifying negative findings from the literature. On the other hand, the way in which the curation protocol was designed, allowed the creation of a corpus to train text mining systems for the detection of negative findings from the literature.

A key point to association class typing, is the availability of corpora to train and validate the RE extraction systems. In this regard, the development of corpora by semi-automatic procedures from databases and text-mined datasets was particularly helpful. Finally, the results on the distinction between drug adverse effects from the therapeutic use of a drug, obtained at the BioCreAtIvE V challenge, showed the value of incorporating background knowledge at the RE step.

5.7 Implications of the choice of document type and document section on the text mining results

It has been suggested by several authors (248–250) that mining full text articles instead of abstracts is preferred, in order to extract all available information from the scientific publications (i.e., information that rarely appears in abstracts, such as experimental measurements). In this thesis, only abstracts from publications were used, for the following reasons: a) abstracts are available for almost all publications in MEDLINE, while this is not the case for the complete article, b) abstracts contain the main findings of the publication, while identifying the relevant information can be more challenging in the full text article and c) abstracts are easy to process, while the full text might be cumbersome, especially for articles that are available only in PDF format.

Interestingly, in the large-scale extraction of gene-disease associations from MEDLINE (Section 0), it was demonstrated that mining only a small fraction of MEDLINE resulted in a large dataset of gene-disease associations. Specifically, from 737,712 abstracts pertaining to human diseases and their genes (approximately 3% of the MEDLINE database), 530,347 associations were obtained between 14,777 genes and 12,650 diseases, which were reported in 355,976 publications. These figures support the notion that the abstracts constitute a rich source of information, at least on the genetic basis of human diseases. Although the potential value of extracting information from the full body of the article should not be underestimated, the issue of quality assessment of these large datasets should be taken into account in the context of database curation pipelines.

Another important aspect in relation extraction involves the span of text considered for the identification of relations between entities. Although some preliminary works based on co-occurrences used entire paragraphs or the abstract, most recent RE systems based on linguistic information, consider the RE problem at the sentence level. Although this approach can result in a more accurate detection of relations by minimizing the *FP*, the *recall* may not be optimal due to the use of co-references and anaphoras in natural language.

The BeFree system was designed to cope with relations stated in a single sentence. This limitation needed to be circumvented for the BioCreAtIvE V challenge, where the relations to be extracted crossed the sentence boundaries. The proposed approach involved the use of patterns and background knowledge (see Section 0). The BeFree CID-based approach did not exhibit the highest performance in the challenge in terms of *F-score*, however, it achieved the best *recall*. The best performing (57.03% of *F-score*) system, presented by Xu et al. (2015) (251), applied two supervised learning approaches (i.e., two SVM-based classifiers trained at the sentence- and document- level, respectively) to extract CID-relations. For the training of their classifiers, basic word-context information (such as bigram of words between the target chemical and disease entities), knowledge features from biomedical databases and document-context information (such as if the chemical or disease names occurred in the title) were employed. Lately, Le et al. (2016) (235) improved their own results obtained at the Bicreative V CID task (+4.13% of *F-score*), by applying a multi-pass sieve co-reference resolution approach.

5.8 Future perspectives

In this thesis, it was shown that the BeFree system is able to extract relevant biomedical information in the context of applied research projects. However, there is still room for improvement regarding the performance of the system.

At the level of NER, the disambiguation step could be, further, improved by the use of linguistic features and statistical methods that remove *false positives*. In addition, the system could be modified in order to identify composite mentions (e.g., “BRCA1/2” or “cardiac and respiratory complications”), as in (97).

Furthermore, novel strategies could be investigated for the detection of gene/disease names. Specifically, a supervised learning approach could be implemented, which could be trained with our novel curated PsyGeNET corpus. It should be also noted that, the latter constitutes an important corpus that can be, also, used to train and improve the BeFree RE approach.

At the level of RE, future work involves the implementation of approaches for co-reference and anaphora resolution and the extraction of information from full text articles. The presented approach, utilizing background knowledge, is valid yet limited by the fact that only known associations can be detected. In addition, it has to be taken into account that in BioCreAtIvE V, a fraction of the relationships obtained at the document level, was not expressed using anaphoras and co-references.

6 Conclusions

[Talking about computers]

"But they are useless.

They can only give you answers"

Pablo Ruiz Picasso (1881-1973)

The main achievements of this thesis are presented below.

(1) The BeFree system was developed as a text mining tool to extract biomedical information from the literature.

(2) A NER approach based on dictionaries and rules has been developed to detect and identify genes and diseases in text.

(3) The ambiguity between gene and disease names has been addressed in order to properly identify these entities.

(4) A RE approach, based on supervised learning, was developed to extract relationships between biomedical entities, by employing shallow and deep syntactic information.

(5) BeFree achieved state-of-the-art performance for the identification of three different types of relationships relevant to the biomedical field: gene-disease, drug-disease and drug-target associations, as shown by the extensive evaluation performed using different gold standards and applications, including a community challenge.

(6) An approach combining the RE module with background knowledge showed acceptable performance for the identification of drug-induced diseases and distinguishing them from drug therapeutic effects.

(7) We addressed the variety of perspectives that can be used to semantically classify a particular relation between two entities.

(8) We have contributed with several annotated corpora and made them publicly available to support the development of text mining tools.

(9) Automatically generated corpora are suitable for the development of text mining tools for biomedical literature.

(10) BeFree was used to extract relevant biomedical information to develop knowledge resources.

7 Appendix

Appendix 1. DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes

DisGeNET is a comprehensive discovery platform designed to address a variety of questions concerning the genetic underpinning of human diseases. DisGeNET contains over 380 000 associations between >16 000 genes and 13 000 diseases, which makes it one of the largest repositories currently available of its kind. DisGeNET integrates expert-curated databases with text-mined data, covers information on Mendelian and complex diseases, and includes data from animal disease models. It features a score based on the supporting evidence to prioritize gene-disease associations. The web interface supports user-friendly data exploration and navigation. DisGeNET data can also be analysed via the DisGeNET Cytoscape plugin, and enriched with the annotations of other plugins of this popular network analysis software suite. Finally, the information contained in DisGeNET can be expanded and complemented using Semantic Web technologies and linked to a variety of resources already present in the Linked Data cloud. Hence, DisGeNET offers one of the most comprehensive collections of human gene-disease associations and a valuable set of tools for investigating the molecular mechanisms underlying diseases of genetic origin, designed to fulfill the needs of different user profiles, including bioinformaticians, biologists and health-care practitioners.

Piñero J, Queralt-Rosinach N, Bravo À, Deu-Pons J, Bauer-Mehren A, Baron M, Sanz F, Furlong LI. *DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes.* Database. 2015 Jan 1; 2015:bav028.

Piñero J, Queralt-Rosinach N, Bravo À, Deu-Pons J, Bauer-Mehren A, Baron M, Sanz F, Furlong LI. DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. Database (Oxford). 2015 Apr 15;2015:bav028. doi: 10.1093/database/bav028

Appendix 2. PsyGeNET: a knowledge platform on psychiatric disorders and their genes

PsyGeNET (Psychiatric disorders and Genes association NETwork) is a knowledge platform for the exploratory analysis of psychiatric diseasesand their associated genes. PsyGeNET is composed of a database and a web interface supporting data search, visualization, filtering and sharing. PsyGeNET integrates information from DisGeNET and data extracted from the literature by text mining, which has been curated by domain experts. It currently contains 2642 associations between 1271 genes and 37 psychiatric disease concepts. In its first release, PsyGeNET is focused on three psychiatric disorders: major depression, alcohol and cocaine use disorders. PsyGeNET represents a comprehensive, open access resource for the analysis of the molecular mechanisms underpinning psychiatric disorders and their comorbidities.

Gutiérrez-Sacristán A, Grosdidier S, Valverde O, Torrens M, Bravo À, Piñero J, Sanz F, Furlong LI. *PsyGeNET: a knowledge platform on psychiatric disorders and their genes.* Bioinformatics. 2015 May; 11:btv301.

Gutiérrez-Sacristán A, Grosdidier S, Valverde O, Torrens M,
Bravo À, Piñero J, Sanz F, Furlong LI. PsyGeNET: a knowledge
platform on psychiatric disorders and their genes.
Bioinformatics. 2015 Sep 15;31(18):3075-7.
doi: 10.1093/bioinformatics/btv301

Appendix 3. A crowdsourcing workflow for extracting chemical-induced disease relations from free text

Relations between chemicals and diseases are one of the most queried biomedical interactions. Although expert manual curation is the standard method for extracting these relations from the literature, it is expensive and impractical to apply to large numbers of documents, and therefore alternative methods are required. We describe here a crowdsourcing workflow for extracting chemical-induced disease relations from free text as part of the BioCreative V Chemical Disease Relation challenge. Five non-expert workers on the CrowdFlower platform were shown each potential chemical-induced disease relation highlighted in the original source text and asked to make binary judgments about whether the text supported the relation. Worker responses were aggregated through voting, and relations receiving four or more votes were predicted as true. On the official evaluation dataset of 500 PubMed abstracts, the crowd attained a 0.505 F-score (0.475 precision, 0.540 recall), with a maximum theoretical recall of 0.751 due to errors with named entity recognition. The total crowdsourcing cost was $1290.67 ($2.58 per abstract) and took a total of 7 h. A qualitative error analysis revealed that 46.66% of sampled errors were due to task limitations and gold standard errors, indicating that performance can still be improved. All code and results are publicly available at https://github.com/SuLab/crowd_cid_relex.

Li TS, Bravo À, Furlong LI, Good BM, Su AI. *A crowdsourcing workflow for extracting chemical-induced disease relations from free text.* Database. 2016 Jan 1; 2016:baw051.

Bravo À, Furlong LI, Good BM, Su AI. A crowdsourcing workflow for extracting chemical-induced disease relations from free text. Database (Oxford). 2016 Apr 17;2016. pii: baw051. doi: 10.1093/database/baw051

Appendix 4. DisGeNET-RDF: harnessing the innovative power of the Semantic Web to explore the genetic basis of diseases

Motivation: DisGeNET-RDF makes available knowledge on the genetic basis of human diseases in the Semantic Web (SW). Gene-disease associations (GDAs) and their provenance metadata are published as human-readable and machine-processable web resources. The information on GDAs included in DisGeNET-RDF is interlinked to other biomedical databases to support the development of bioinformatics approaches for translational research through evidence-based exploitation of a rich and fully interconnected Linked Open Data (LOD).

Queralt-Rosinach N, Piñero J, Bravo À, Sanz F, Furlong LI. *DisGeNET-RDF: harnessing the innovative power of the Semantic Web to explore the genetic basis of diseases.* Bioinformatics. 2016 Apr 22; btw214.

Queralt-Rosinach N, Piñero J, Bravo À, Sanz F, Furlong LI. DisGeNET-RDF: harnessing the innovative power of the Semantic Web to explore the genetic basis of diseases. Bioinformatics. 2016 Jul 15;32(14):2236-8. doi:10.1093/bioinformatics/btw214.

Bibliography

1. Chen H, Fuller SS, Friedman C. Knowledge Management, Data Mining, and Text Mining in Medical Informatics. Medical Informatics: Knowledge Management and Data Mining in Biomedicine. 2005. p. 3–33.

2. Huang C, Lu Z. Community challenges in biomedical text mining over 10 years: success , failure and the future. Briefings in bioinformatics. 2016. p. 132–44.

3. MEDLINE [Internet]. Available from: https://www.nlm.nih.gov/bsd/pmresources.html

4. Holzinger A, Dehmer M, Jurisica I. Knowledge Discovery and interactive Data Mining in Bioinformatics - State-of-the-Art, future challenges and research directions. BMC Bioinformatics. 2014;15 Suppl 6:I1.

5. Rebholz-Schuhmann D, Oellrich A, Hoehndorf R. Text-mining solutions for biomedical research: enabling integrative biology. Nat Rev Genet. Nature Publishing Group; 2012;13(12):829–39.

6. The Seventh IEEE Conference on Artificial Intelligence Application [Internet]. Available from: http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=331

7. Grishman R, Sundheim B. Message Understanding Conference-6: A Brief History. Proceedings of the 16th conference on Computational linguistics. 1996. p. 466–71.

8. Verspoor K, Cohen KB, Goertzel B, Mani I. BioNLP'06 Linking Natural Language Processing and Biology: Towards Deeper Biological Literature Analysis. Proceedings of the HLT-NAACL BioNLP Workshop on Linking Natural Language and Biology. New York, NY; 2006.

9. Lindsey WT, Olin BR. PubMed Searches: Overview and Strategies for Clinicians. Nutr Clin Pract. 2013;28(2):165–76.

10. Scopus [Internet]. Available from: https://www.elsevier.com/solutions/scopus

11. Google Scholar [Internet]. Available from: https://scholar.google.es/intl/en/scholar/about.html

12. Davis a. P, Grondin CJ, Lennon-Hopkins K, Saraceni-Richards C, Sciaky D, King BL, et al. The Comparative Toxicogenomics Database's 10th year anniversary: update 2015. Nucleic Acids Res. 2014;43(D1):D914–20.

13. Kuhn M, Letunic I, Jensen LJ, Bork P. The SIDER database of drugs and side effects. Nucleic Acids Res. 2015;1–5.

14. Whirl-Carrillo M, McDonagh EM, Hebert JM, Gong L, Sangkuhl K, Thorn CF, et al. Pharmacogenomics knowledge for personalized medicine. Clin Pharmacol Ther. 2012;92(4):414–7.

15. Chapman WW, Cohen KB. Current issues in biomedical text mining and natural language processing. J Biomed Inform. 2009;42(5):757–9.

16. NCBI Gene [Internet]. Available from: http://www.ncbi.nlm.nih.gov/gene

17. Hirschman L, Yeh A, Blaschke C, Valencia A. Overview of BioCreAtIvE: critical assessment of information extraction for biology. BMC Bioinformatics. 2005;6(1):S1.

18. Krallinger M, Morgan A, Smith L, Leitner F, Tanabe L, Wilbur J, et al. Evaluation of text-mining systems for biology: overview of the Second BioCreative community challenge. Genome Biol. 2008;9(2):S1.

19. Arighi CN, Roberts PM, Agarwal S, Bhattacharya S, Cesareni G, Chatr-Aryamontri A, et al. BioCreative III interactive task: an overview. BMC Bioinformatics. 2011;12(8):S1.

20. Kim J-D, Ohta T, Tsuruoka Y, Tateisi Y, Collier N. Introduction to the Bio-entity Recognition Task at JNLPBA. Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and Its Applications. 2004. p. 70–5.

21. Rebholz-Schuhmann D, Yepes AJ, Li C, Kafkas S, Lewin I, Kang N, et al. Assessment of NER solutions against the first and second CALBC Silver Standard Corpus. CEUR Workshop Proceedings. 2010. p. 63–71.

22. Rebholz-Schuhmann D, Yepes AJJJ, Van Mulligen EM, Kang N, Kors J, Milward D, et al. CALBC silver standard corpus. J Bioinform Comput Biol. 2010;8(1):163–79.

23. Kim J-D, Ohta T, Pyysalo S, Kano Y, Tsujii J. Overview of BioNLP'09 shared task on event extraction. Proceedings of the Workshop on BioNLP Shared Task - BioNLP '09. 2009. p. 1.

24. Kim, Wang Y, Takagi, Yonezawa. Overview of Genia Event Task in BioNLP Shared Task 2011. In Proceedings of the BioNLP Shared Task 2011 Workshop 2011 Association for Computational Linguistics. 2011. p. 7–15.

25. Kim J-D, Wang Y, Yasunori Y. The Genia Event Extraction Shared Task, 2013 Edition-Overview. BioNLP Shared Task 2013 Workshop. 2013. p. 8–15.

26. Nédellec C. Learning language in logic-genic interaction extraction challenge. Proceedings of the 4th Learning Language in Logic Workshop (LLL05). 2005. p. vol.7 1–7.

27. Fleuren WWM, Alkema W. Application of text mining in the biomedical domain. Methods. 2015;74:97–106.

28. NLM Unified Medical Language System (UMLS) [Internet]. Available from: http://www.nlm.nih.gov/research/umls/

29. Özgür A, Vu T, Erkan G, Radev DR. Identifying gene-disease associations using centrality on a literature mined gene-interaction network. Bioinformatics. 2008;24(13).

30. Bundschus M, Dejori M, Stetter M, Tresp V, Kriegel H-P. Extraction of semantic biomedical relations from text using conditional random fields. BMC Bioinformatics. 2008;9:207.

31. Kim J, Kim H, Yoon Y, Park S. LGscore: A method to identify disease-related genes using biological literature and Google data. J Biomed Inform. 2015;54:270–82.

32. Chen H, Sharp BM. Content-rich biological network constructed by mining PubMed abstracts. BMC Bioinformatics. 2004;5:147.

33. Hoffmann R, Valencia A. Implementing the iHOP concept for navigation of biomedical literature. Bioinformatics. 2005;21(S2).

34. Navlakha S, Kingsford C. The power of protein interaction networks for associating genes with diseases. Bioinformatics. 2010;26(8):1057–63.

35. Hristovski D, Peterlin B, Mitchell JA, Humphrey SM. Using literature-based discovery to identify disease candidate genes. Int J Med Inform. 2005;74(2-4):289–98.

36. Wikipedia: Precision and Recall [Internet]. Available from: https://en.wikipedia.org/wiki/Precision_and_recall

37. Islamaj Dogan R, Murray GC, Névéol A, Lu Z. Understanding PubMed® user search behavior through log analysis. 2009;2009:18.

38. Pelat C, Turbelin C, Bar-Hen A, Flahault A, Valleron AJ. More diseases tracked by using google trends. Emerging Infectious Diseases. 2009. p. 1327–8.

39. Doğan RI, Lu Z. An improved corpus of disease mentions in PubMed citations. Proceedings of the 2012 Workshop on Biomedical Natural Language Processing. 2012. p. 91–9.

40. Chun H-W, Tsuruoka Y, Kim J-D, Shiba R, Nagata N, Hishiki T, et al. Extraction of gene-disease relations from Medline using domain dictionaries and machine learning. Pac Symp Biocomput. 2006;4–15.

41. Hakenberg J, Voronov D, Nguyen VH, Liang S, Anwar SS, Leaman R, et al. A SNPshot of PubMed to find associations between genetic variants, drugs, and diseases. J Biomed Inform. 2012;

42. Liu K, Hogan WR, Crowley RS. Natural Language Processing methods and systems for biomedical ontology learning. J Biomed Inform. 2011;44(1):163–79.

43. Díaz NPC, López MMM. An Analysis of Biomedical Tokenization : Problems and Strategies. Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis. Lisbon; 2015. p. 40–9.

44. Jensen LJ, Saric J, Bork P. Literature mining for the biologist: from information retrieval to biological discovery. Nat Rev Genet. 2006;7(2):119–29.

45. Clegg AB, Shepherd AJ. Benchmarking natural-language parsers for biological applications using dependency graphs. BMC Bioinformatics. 2007;8:24.

46. Penn Treebank Tag Set [Internet]. Available from: http://www.comp.leeds.ac.uk/ccalas/tagsets/upenn.html

47. GENIA Tagger [Internet]. Available from: http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/

48. Illinois Shallow Parser [Internet]. Available from: http://cogcomp.cs.illinois.edu/page/demo_view/ShallowParse

49. Apache OpenNLP [Internet]. Available from: http://opennlp.apache.org/index.html

50. Lease M, Charniak E. Parsing biomedical literature. Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCNLP '05). Jeju Island, Korea; 2005. p. 58–69.

51. Pyysalo S, Salakoski T, Aubin S, Nazarenko A. Lexical adaptation of link grammar to the biomedical sublanguage: A comparative evaluation of three approaches. BMC Bioinformatics,. 2006;7.

52. Rimell L, Clark S. Porting a lexicalized-grammar parser to the biomedical domain. J Biomed Inform. 2009;42(5):852–65.

53. Stanford Lexical Parser [Internet]. Available from: http://nlp.stanford.edu/software/lex-parser.shtml

54. Link Parser [Internet]. Available from: http://www.link.cs.cmu.edu/link/

55. Enju Parser [Internet]. Available from: http://www.nactem.ac.uk/enju/

56. Genia Dependency Parser [Internet]. Available from: http://sagae.bitbucket.org/gdep/

57. Bikel Parser [Internet]. Available from: http://www.cis.upenn.edu/~dbikel/#stat-parser

58. Rodriguez-Esteban R. Biomedical text mining and its applications. PLoS Comput Biol. 2009;5.

59. Nenadic G, Spasic I, Ananiadou S. Mining biomedical abstracts: What's in a term? International Conference on Natural Language Processing. Springer Berlin Heidelberg; 2004. p. 797–806.

60. Heyninck K, De Valck D, Vanden Berghe W, Van Criekinge W, Contreras R, Fiers W, et al. The zinc finger protein A20 inhibits TNF-induced NF-kappaB-

dependent gene expression by interfering with an RIP- or TRAF2-mediated transactivation signal and directly binds to a novel NF-kappaB-inhibiting protein ABIN. J Cell Biol. 1999;145(7):1471–82.

61. Chen L, Liu H, Friedman C. Gene name ambiguity of eukaryotic nomenclatures. Bioinformatics. 2005;21(2):248–56.

62. UniProt [Internet]. Available from: http://www.uniprot.org/

63. Ananiadou S, McNaught J. Text Mining for Biology and Biomedicine. Computational Linguistics. 2006. 135-140 p.

64. Li L, Zhou R, Huang D. Two-phase biomedical named entity recognition using CRFs. Comput Biol Chem. 2009;33(4):334–8.

65. Khoo C, Chan S, Niu Y. Extracting causal knowledge from a medical database using graphical patterns. Proc 38th Annual Meeting on Association for Computational Linguistics. 2000. p. 336–43.

66. Fukuda K, Tamura A, Tsunoda T, Takagi T. Toward information extraction: identifying protein names from biological papers. Pac Symp Biocomput. 1998;707–18.

67. Tanabe L, Wilbur WJ. Tagging gene and protein names in biomedical text. Bioinformatics. 2002;18(8):1124–32.

68. Hakenberg J, Bickel S, Plake C, Brefeld U, Zahn H, Faulstich L, et al. Systematic feature evaluation for gene name recognition. BMC Bioinformatics. 2005;6 Suppl 1:S9.

69. Chang JT, Schütze H, Altman RB. GAPSCORE: Finding gene and protein names one word at a time. Bioinformatics. 2004;20(2):216–25.

70. Zhou G, Su J. Exploring Deep Knowledge Resources in Biomedical Name Recognition. Workshop on Natural Language Processing in Biomedicine and Its Applications at COLING. 2004. p. 96–9.

71. Collier N, Nobata C, Tsujii J. Extracting the names of genes and gene products with a hidden Markov model. Proceedings of the 18th conference on Computational linguistics-Volume 1. 2000. p. 201–7.

72. Corbett P, Copestake A. Cascaded classifiers for confidence-based chemical named entity recognition. BMC Bioinformatics. 2008;9 Suppl 11:S4.

73. Saha SK, Sarkar S, Mitra P. Feature selection techniques for maximum entropy based biomedical named entity recognition. J Biomed Inform. 2009;42(5):905–11.

74. Li Y, Lin H, Yang Z. Incorporating rich background knowledge for gene named entity classification and recognition. BMC Bioinformatics. 2009;10:223.

75. Wang H, Zhao T, Tan H, Zhang S. Biomedical Named Entity Recognition Based on Classifiers Ensemble. Int J Comput Sci Appl. 2008;5(2):1–11.

76. Mccallum A, Freitag D, Pereira F. Maximum Entropy Markov Models for Information Extraction and Segmentation. ICML. 2000. p. 591–8.

77. Neves ML, Carazo J-M, Pascual-Montano A. Moara: a Java library for extracting and normalizing gene and protein mentions. BMC Bioinformatics. 2010;11:157.

78. Hanisch D, Fundel K, Mevissen H-T, Zimmer R, Fluck J. ProMiner: rule-based protein and gene entity recognition. BMC Bioinformatics. 2005;6 Suppl 1:S14.

79. Sasaki Y, Tsuruoka Y, McNaught J, Ananiadou S. How to make the most of NE dictionaries in statistical NER. BMC Bioinformatics. 2008;9 Suppl 11:S5.

80. Proux D, Rechenmann F, Julliard L, Pillet V, Jacq B. Detecting gene symbols and names in biological texts: a first step toward pertinent information extraction. Genome Inf Ser Work Genome Inform. 1998;9:72–80.

81. Andrade MA, Valencia A. Automatic extraction of keywords from scientific text: Application to the knowledge domain of protein families. Bioinformatics. 1998;14(7):600–7.

82. Katherine A. Guidelines for Formatting Gene and Protein Names [Internet]. [cited 2015 Aug 5]. Available from: http://www.biosciencewriters.com/Guidelines-for-Formatting-Gene-and-Protein-Names.aspx

83. Shows TB, Alper CA, Bootsma D, Dorf M, Douglas T, Huisman T, et al. International system for human gene nomenclature (1979) ISGN (1979). Cytogenet Cell Genet. SWITZERLAND; 1979;25(1-4):96–116.

84. Franzén K, Eriksson G, Olsson F, Asker L, Lidén P, Cöster J. Protein names and how to find them. International Journal of Medical Informatics. 2002. p. 49–61.

85. Leaman R, Gonzalez G. BANNER: an executable survey of advances in biomedical named entity recognition. Pac Symp Biocomput. 2008;652–63.

86. Yeh A, Morgan A, Colosimo M, Hirschman L. BioCreAtIvE task 1A: gene mention finding evaluation. BMC Bioinformatics. 2005;6 Suppl 1:S2.

87. Hirschman L, Colosimo M, Morgan A, Yeh A. Overview of BioCreAtIvE task 1B: normalized gene lists. BMC Bioinformatics. 2005;6 Suppl 1:S11.

88. Morgan AA, Lu Z, Wang X, Cohen AM, Fluck J, Ruch P, et al. Overview of BioCreative II gene normalization. Genome Biol. 2008;9 Suppl 2:S3.

89. Settles B. ABNER: An open source tool for automatically tagging genes, proteins and other entity names in text. Bioinformatics. 2005;21(14):3191–2.

90. Hakenberg J, Royer L, Plake C, Strobelt H, Schroeder M. Me and my friends: gene mention normalization with background knowledge. Proc 2nd BioCreative Challenge Evaluation Workshop. 2007. p. 141–4.

91. Fundel K, Zimmer R. Human Gene Normalization by an Integrated Approach including Abbreviation Resolution and Disambiguation. Second BioCreative Challenge Evaluation Workshop. 2007.

92. Fluck J, Mevissen HT, Dach H, Oster M, Hofmann-apitius M. ProMiner: Recognition of Human Gene and Protein Names using regularly updated Dictionaries. Proceedings of the Second BioCreative Challenge Evaluation Workshop Centro Nacional de Investigaciones Oncologicas, CNIO, Madrid, Spain, 2007. 2007. p. 149–51.

93. Hakenberg J, Plake C, Leaman R, Schroeder M, Gonzalez G. Inter-species normalization of gene mentions with GNAT. Bioinformatics. 2008;24(16).

94. Lau WW, Johnson C a, Becker KG. Rule-based human gene normalization in biomedical text with confidence estimation. Comput Syst Bioinformatics Conf. 2007;6:371–9.

95. Lu Z, Kao H-Y, Wei C-H, Huang M, Liu J, Kuo C-J, et al. The gene normalization task in BioCreative III. BMC Bioinformatics. 2011;12 Suppl 8(8):S2.

96. Li L, Liu S, Li L, Fan W, Huang D, Zhou H. A multistage gene normalization system integrating multiple effective methods. PLoS One. 2013;8(12).

97. Wei CH, Kao HY, Lu Z. GNormPlus: An Integrative Approach for Tagging Genes, Gene Families, and Protein Domains. Biomed Res Int. 2015;2015.

98. Li L, Fan W, Huang D, Dang Y, Sun J. Boosting performance of gene mention tagging system by hybrid methods. J Biomed Inform. 2012;45(1):156–64.

99. Wei C-H, Kao H-Y. Cross-species gene normalization by species inference. BMC Bioinformatics. 2011;12(Suppl 8):S5.

100. Wei CH, Huang IC, Hsu YY, Kao HY. Normalizing biomedical name entities by similarity-based inference network and de-ambiguity mining. Proceedings

of the 2009 9th IEEE International Conference on Bioinformatics and BioEngineering, BIBE 2009. 2009. p. 461–6.

101. Dogan R, Lu Z. An inference method for disease name normalization. AAAI Fall Symposium - Technical Report. 2012. p. 8–13.

102. Jimeno A, Jimenez-Ruiz E, Lee V, Gaudan S, Berlanga R, Rebholz-Schuhmann D. Assessment of disease named entity recognition on a corpus of annotated sentences. BMC Bioinformatics. 2008;9 Suppl 3:S3.

103. Gurulingappa H, Klinger R, Hofmann-apitius M, J. An Empirical Evaluation of Resources for the Identification of Diseases and Adverse Effects in Biomedical Literature. 2nd Workshop on Building and Evaluating Resources for Biomedical Text Mining. 2010. p. 15–22.

104. Leaman R, Doğan RI, Lu Z. DNorm: Disease name normalization with pairwise learning to rank. Bioinformatics. 2013;29(22):2909–17.

105. NLM Medical Subject Headings (MeSH) [Internet]. Available from: http://www.ncbi.nlm.nih.gov/mesh

106. Systematized Nomenclature of Medicine – Clinical Terms (Snomed CT) [Internet]. Available from: http://www.ihtsdo.org/snomed-ct

107. International Classification of Diseases (ICD) [Internet]. Available from: http://www.who.int/classifications/icd/en/

108. Online Mendelian Inheritance in Man (OMIM) [Internet]. Available from: http://www.omim.org

109. Uzuner O, South BR, Shen S, Duvall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. J Am Med Inform Assoc. 2011;18(5):552–6.

110. Suominen H, Salanterä S, Velupillai S, Chapman WW, Savova G, Elhadad N, et al. Overview of the ShARe/CLEF eHealth evaluation lab 2013. Lect Notes

Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics). 2013;8138 LNCS:212–31.

111. Pradhan S, Manandhar S, Savova G. SemEval-2014 Task 7 : Analysis of Clinical Text. SemEval. 2014;199.99:54–62.

112. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform. 2008;128–44.

113. Wu ST-I, Liu H. Semantic characteristics of NLP-extracted concepts in clinical notes vs. biomedical literature. Proceedings of AMIA. 2010. p. 1550–8.

114. Leaman R, Miller C. Enabling Recognition of Diseases in Biomedical Text with Machine Learning : Corpus and Benchmark. Proceedings of the 3rd International Symposium on Languages in Biology and Medicine (LBM). 2009. p. 82–9.

115. Carpenter B. Phrasal Queries with LingPipe and Lucene : Ad Hoc Genomics Text Retrieval. TREC. 2004;1:1–10.

116. Kang N, Singh B, Afzal Z, van Mulligen EM, Kors JA. Using rule-based natural language processing to improve disease normalization in biomedical text. Journal of the American Medical Informatics Association. 2012.

117. Chowdhury M, Faisal M. Disease mention recognition with specific features. Proceedings of the 2010 Workshop on …. 2010. p. 83–90.

118. Hahn U, Buyko E, Landefeld R, Poprat M, Tomanek K, Wermter J. An Overview of JCORE, the JULIE Lab UIMA Component Repository. Language Resources and Evaluation. 2008. p. 1–7.

119. Zhong H, Hu X. Disease Named Entity Recognition by Machine Learning Using Semantic Type of Metathesaurus. Int J Mach Learn Comput. 2013;3(6):494–8.

120. Aronson A. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proceedings / AMIA Annual Symposium. 2001. p. 17–21.

121. Schuemie MJ, Jelier R, Kors JA. Peregrine: Lightweight gene name normalization by dictionary lookup. Proc of the Second BioCreative Challenge Evaluation Workshop. 2007. p. 131–3.

122. Doğan RI, Leaman R, Lu Z. NCBI disease corpus: A resource for disease name recognition and concept normalization. J Biomed Inform. 2014;47:1–10.

123. Neveol, Aurelie, Lu, Zhiyong. Automatic integration of drug indications from multiple health resources. 1st ACM International Health Informatics Symposium, IHI'10, November 11, 2010 - November 12, 2010. 2010. p. 666–73.

124. Sampathkumar H, Chen X, Luo B. Mining adverse drug reactions from online healthcare forums using hidden Markov model. BMC Med Inform Decis Mak. 2014;14(1):91.

125. Xu R, Wang Q. Automatic construction of a large-scale and accurate drug-side-effect association knowledge base from biomedical literature. J Biomed Inform. 2014;51:191–9.

126. Gurulingappa H, Mateen-Rajput A, Toldo L. Extraction of potential adverse drug events from medical case reports. J Biomed Semantics. 2012;3:15.

127. Hakenberg J, Plake C, Leser U, Kirsch H, Rebholz-schuhmann D. LLL ' 05 Challenge : Genic Interaction Extraction - Identification of Language Patterns Based on Alignment and Finite State Automata. 2005;

128. Krallinger M, Leitner F, Rodriguez-Penagos C, Valencia A. Overview of the protein-protein interaction annotation extraction task of BioCreative II. Genome Biol. 2008;9 Suppl 2(Suppl 2):S4.

129. Simpson MS, Demner-Fushman D. Biomedical Text Mining: A Survey of Recent Progress. Mining Text Data. 2012. p. 465–517.

130. Wei C-H, Peng Y, Leaman R, Davis AP, Mattingly CJ, Li J, et al. Overview of the BioCreative V Chemical Disease Relation (CDR) Task. Proc Fifth BioCreative Chall Eval Work. 2015;154–66.

131. Cohen KB, Hunter L. Getting started in text mining. PLoS Computational Biology. 2008. p. 0001–3.

132. Faro A, Giordano D, Spampinato C. Combining literature text mining with microarray data: Advances for system biology modeling. Brief Bioinform. 2012;13(1):61–82.

133. Jenssen TK, Laegreid A, Komorowski J, Hovig E. A literature network of human genes for high-throughput analysis of gene expression. Nat Genet. 2001;28(1):21–8.

134. Wang X, Hripcsak G, Markatou M, Friedman C. Active Computerized Pharmacovigilance Using Natural Language Processing, Statistics, and Electronic Health Records: A Feasibility Study. J Am Med Informatics Assoc. 2009;16(3):328–37.

135. Craven M, Kumlien J. Constructing biological knowledge bases by extracting information from text sources. Proc Int Conf Intell Syst Mol Biol. 1999;77–86.

136. Pletscher-Frankild S, Pallejà A, Tsafou K, Binder JX, Jensen LJ. DISEASES: Text mining and data integration of disease-gene associations. Methods. 2015;74:83–9.

137. Kandula S, Zeng-Treitler Q. Exploring relations among semantic groups: A comparison of concept co-occurrence in biomedical sources. Studies in Health Technology and Informatics. 2010. p. 995–9.

138. Alako BTF, Veldhoven A, van Baal S, Jelier R, Verhoeven S, Rullmann T, et al. CoPub Mapper: mining MEDLINE based on search term co-publication. BMC Bioinformatics. 2005;6(1):51.

139. Wren JD, Garner HR. Shared relationship analysis: Ranking set cohesion and commonalities within a literature-derived relationship network. Bioinformatics. 2004;20(2):191–8.

140. Pyysalo S, Airola A, Heimonen J, Björne J, Ginter F, Salakoski T. Comparative analysis of five protein-protein interaction corpora. BMC Bioinformatics. 2008;9 Suppl 3:S6.

141. Airola A, Pyysalo S, Björne J, Pahikkala T, Ginter F, Salakoski T. All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning. BMC Bioinformatics. 2008;9 Suppl 11(S2):1.

142. Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: Interpreting hypernymic propositions in biomedical text. J Biomed Inform. 2003;36(6):462–77.

143. Friedman C, Liu H, Shagina L, Johnson S, Hripcsak G. Evaluating the UMLS as a source of lexical knowledge for medical language processing. Proceedings / AMIA Annual Symposium. 2001. p. 189–93.

144. Fundel K, Küffner R, Zimmer R. RelEx - Relation extraction using dependency parse trees. Bioinformatics. 2007;23(3):365–71.

145. Divoli A, Attwood TK. BiolE: Extracting informative sentences from the biomedical literature. Bioinformatics. 2005;21(9):2138–9.

146. Šarić J, Jensen LJ, Ouzounova R, Rojas I, Bork P. Extraction of regulatory gene/protein networks from Medline. Bioinformatics. 2006;22(6):645–50.

147. Thomas J, Milward D. Automatic Extraction of Protein Interactions from Scientific Abstracts. Pacific Symposium on Biocomputing. 2000. p. 538–49.

148. Yakushiji A, Tateisi Y, Miyao Y, Tsujii J. Event extraction from biomedical papers using a full parser. Pacific Symposium on Biocomputing. 2001. p. 408–19.

149. Hakenberg J, Leaman R, Ha Vo N, Jonnalagadda S, Sullivan R, Miller C, et al. Efficient extraction of protein-protein interactions from full-text articles. IEEE/ACM Trans Comput Biol Bioinforma. 2010;7(3):481–94.

150. Hao Y, Zhu X, Huang M, Li M. Discovering patterns to extract protein-protein interactions from the literature: Part II. Bioinformatics. 2005;21(15):3294–300.

151. Caporaso JG, Baumgartner WA, Randolph DA, Cohen KB, Hunter L. Rapid pattern development for concept recognition systems: application to point mutations. J Bioinform Comput Biol. 2007;5(6):1233–59.

152. Kim S, Yoon J, Yang J, Park S. Walk-weighted subsequence kernels for protein-protein interaction extraction. BMC Bioinformatics. 2010;11:107.

153. Zhang Y, Lin H, Yang Z, Wang J, Li Y. A Single Kernel-Based Approach to Extract Drug-Drug Interactions from Biomedical Literature. PLoS One. 2012;7(11).

154. Liu H, Hunter L, Kešelj V, Verspoor K. Approximate Subgraph Matching-Based Literature Mining for Biomedical Events and Relations. PLoS One. 2013;8(4).

155. McClosky D, Riedel S, Surdeanu M, McCallum A, Manning CD. Combining joint models for biomedical event extraction. BMC Bioinformatics. 2012. p. 1.

156. Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. J Am Med Informatics Assoc. 2011;18(5):544–51.

157. Giuliano C, Lavelli A, Romano L. Exploiting Shallow Linguistic Information for Relation Extraction from Biomedical Literature. Proc of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics (EACL2006). European Chapter of the Association for Computational Linguistics; 2006. p. 401–8.

158. Miwa M, Sætre R, Miyao Y, Tsujii J. Protein-protein interaction extraction by leveraging multiple kernels and parsers. Int J Med Inform. 2009;78(12).

159. Yang Z, Tang N, Zhang X, Lin H, Li Y, Yang Z. Multiple kernel learning in protein-protein interaction extraction from biomedical literature. Artif Intell Med. 2011;51(3):163–73.

160. Segura-Bedmar I, Martínez P, de Pablo-Sánchez C. Using a shallow linguistic kernel for drug-drug interaction extraction. J Biomed Inform. 2011;44(5):789–804.

161. Bunescu RC, Mooney RJ. Subsequence kernels for relation extraction. Adv Neural Inf Process Syst. 2006;18:171.

162. Zelenko D, Aone C, Richardella A. Kernel Methods for Relation Extraction. J Mach Learn Res. 2003;3(6):1083–106.

163. Chowdhury MFM, Lavelli A. Combining Tree Structures, Flat Features and Patterns for Biomedical Relation Extraction. Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. 2012. p. 420–9.

164. Culotta A, Sorensen J. Dependency tree kernels for relation extraction. Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics ACL 04. 2004. p. 423 – es.

165. Giuliano C, Lavelli A, Romano L. Relation extraction and the influence of automatic named-entity recognition. ACM Trans Speech Lang Process. 2007;5(1):1–26.

166. Hahn U, Bretonnel Cohen K, Garten Y, Shah NH. Mining the pharmacogenomics literature-A survey of the state of the art. Brief Bioinform. 2012;13(4):460–94.

167. Buyko E, Beisswanger E, Hahn U. The extraction of pharmacogenetic and pharmacogenomic relations--a case study using PharmGKB. Pac Symp Biocomput. 2012;376–87.

168. Pakhomov S, McInnes BT, Lamba J, Liu Y, Melton GB, Ghodke Y, et al. Using PharmGKB to train text mining approaches for identifying potential gene targets for pharmacogenomic studies. J Biomed Inform. 2012;45(5):862–9.

169. Xu R, Wang Q. A knowledge-driven conditional approach to extract pharmacogenomics specific drug-gene relationships from free text. J Biomed Inform. 2012;45(5):827–34.

170. Percha B, Garten Y, Altman RB. Discovery and explanation of drug-drug interactions via text mining. Pac Symp Biocomput. 2012;(DDI):410–21.

171. Kim S, Liu H, Yeganova L, Wilbur WJ. Extracting drug-drug interactions from literature using a rich feature-based linear kernel approach. J Biomed Inform. 2015;55:23–30.

172. Bui QC, Sloot PMA, Van Mulligen EM, Kors JA. A novel feature-based approach to extract drug-drug interactions from biomedical text. Bioinformatics. 2014;30(23):3365–71.

173. Fiszman M, Demner-Fushman D, Kilicoglu H, Rindflesch TC. Automatic summarization of MEDLINE citations for evidence-based medical treatment: A topic-oriented evaluation. J Biomed Inform. 2009;42(5):801–13.

174. Kang N, Singh B, Bui C, Afzal Z, van Mulligen EM, Kors JA. Knowledge-based extraction of adverse drug events from biomedical text. BMC Bioinformatics. 2014;15(1):64.

175. Baker NC, Hemminger BM. Mining connections between chemicals, proteins, and diseases extracted from Medline annotations. J Biomed Inform. 2010;43(4):510–9.

176. Frijters R, van Vugt M, Smeets R, van Schaik R, de Vlieg J, Alkema W. Literature mining for the discovery of hidden connections between drugs, genes and diseases. PLoS Comput Biol. 2010;6(9).

177. Zhang M, Zhang J, Su J. Exploring Syntactic Features for Relation Extraction using a Convolution Tree Kernel. Proceedings of the Human Language Technology Conference of the NAACL, Main Conference. 2006.

178. Zhao S, Grishman R. Extracting relations with integrated information using kernel methods. Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics ACL 05. 2005. p. 419–26.

179. Chowdhury M, Lavelli A. An Evaluation of the Effect of Automatic Preprocessing on Syntactic Parsing for Biomedical Relation Extraction. LREC. 2012;544–51.

180. Hou WJ, Chen HY. Rule extraction in gene-disease relationship discovery. Gene. 2013;518(1):132–8.

181. Krallinger M, Leitner F, Valencia A. Analysis of biological processes and diseases using text mining approaches. Methods Mol Biol. 2010;593:341–82.

182. Adamic LA, Wilkinson D, Huberman BA, Adar E. A literature based method for identifying gene-disease connections. Proceedings - IEEE Computer Society Bioinformatics Conference, CSB 2002. 2002. p. 109–17.

183. Al-Mubaid H, Singh RK. A new text mining approach for finding protein-to-disease associations. Am J Biochem Biotechnol. 2005;1(3):145.

184. Tsuruoka Y, Tsujii J, Ananiadou S. FACTA: A text search engine for finding associated biomedical concepts. Bioinformatics. 2008;24(21):2559–60.

185. Chun HW, Tsuruoka Y, Kim JD, Shiba R, Nagata N, Hishiki T, et al. Automatic recognition of topic-classified relations between prostate cancer and genes using MEDLINE abstracts. CEUR Workshop Proceedings. 2006. p. 1.

186. Buyko E, Faessler E, Wermter J, Hahn U. Event extraction from trimmed dependency graphs. Proceedings of the Workshop on BioNLP Shared Task BioNLP 09. 2009. p. 19.

187. Mork S, Pletscher-Frankild S, Caro AP, Gorodkin J, Jensen LJ. Protein-driven inference of miRNA-disease associations. Bioinformatics. 2014;30(3):392–7.

188. Xu D, Zhang M, Xie Y, Wang F, Chen M, Zhu KQ, et al. DTMiner: Identification of potential disease targets through biomedical literature mining. Bioinformatics. 2016;btw503.

189. Verspoor K, Yepes AJ, Cavedon L, McIntosh T, Herten-Crabb A, Thomas Z, et al. Annotating the biomedical literature for the human variome. Database. 2013;2013.

190. Neves M, Damaschun A, Kurtz A, Leser U. Annotating and evaluating text for stem cell research. Proc Third Work Build Eval Resour Biomed Text Min (BioTxtM 2012) Lang Resour Eval 2012. 2012;16–23.

191. Van Mulligen EM, Fourrier-Reglat A, Gurwitz D, Molokhia M, Nieto A, Trifiro G, et al. The EU-ADR corpus: Annotated drugs, diseases, targets, and their relationships. J Biomed Inform. 2012;45(5):879–84.

192. Kim JD, Ohta T, Tateisi Y, Tsujii J. GENIA corpus - A semantically annotated corpus for bio-textmining. Bioinformatics. 2003;19(S1).

193. Smith L, Tanabe LK, Ando RJ nee, Kuo C-J, Chung I-F, Hsu C-N, et al. Overview of BioCreative II gene mention recognition. Genome Biol. 2008;9 Suppl 2:S2.

194. Krallinger M, Leitner F, Rabal O. Overview of the chemical compound and drug name recognition (CHEMDNER) task. Proceedings of the Fourth BioCreative Challenge Evaluation Workshop. 2013. p. 2–33.

195. Herrero-Zazo M, Segura-Bedmar I, Martínez P, Declerck T. The DDI corpus: An annotated corpus with pharmacological substances and drug-drug interactions. J Biomed Inform. 2013;46(5):914–20.

196. Neves M. An analysis on the entity annotations in biological corpora. F1000Research. 2014;3:96.

197. Segura-Bedmar I, Martínez P, Sánchez-Cisneros D. The 1st DDIExtraction-2011 challenge task: Extraction of Drug-Drug Interactions from biomedical texts. CEUR Workshop Proceedings. 2011. p. 1–9.

198. Collier N, Park HS, Ogata N, Tateishi Y, Nobata C, Ohta T, et al. The GENIA project: corpus-based knowledge acquisition and information extraction from genome research papers. Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics. 1999. p. 271–2.

199. Kazama J, Makino T, Ohta Y, Tsujii J. Tuning Support Vector Machines for Biomedical Named Entity Recognition. Proceedings of the ACL02 workshop on Natural language processing in the biomedical domain. 2002. p. 1–8.

200. Shen D, Zhang J, Zhou G, Su J, Tan C-L. Effective adaptation of a Hidden Markov Model-based named entity recognizer for biomedical domain. Proceedings of the ACL 2003 workshop on Natural language processing in biomedicine -. 2003. p. 49–56.

201. Tanabe L, Xie N, Thom LH, Matten W, Wilbur WJ. GENETAG: a tagged corpus for gene/protein named entity recognition. BMC Bioinformatics. 2005;6 Suppl 1:S3.

202. Rosario B, Hearst MA. Classifying semantic relations in bioscience texts. Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. 2004. p. 430.

203. Kulick S, Bies A, Liberman M, Mandel M, Mcdonald R, Palmer M, et al. Integrated Annotation for Biomedical Information Extraction. HLT-NAACL Workshop: Boilink 2004 Linking Biological Literature, Ontologies and Databases. 2004. p. 61–8.

204. Arizona Disease Corpus Annotation Guidelines [Internet]. Available from: http://diego.asu.edu/downloads/AZDCAnnotationGuidelines_v013.pdf

205. Thomas P, Rockt T, Leser U, Street G, Wce L. SETH detects and normalizes genetic variants in text. Bioinformatics. 2016;2–4.

206. Leaman R, Wei C, Lu Z. tmChem : a high performance approach for chemical named entity recognition and normalization. J Cheminform. Chemistry Central Ltd; 2015;7(Suppl 1):S3.

207. Python 2.7.0 release [Internet]. Available from: https://www.python.org/download/releases/2.7/

208. Loper E, Bird S. NLTK: the Natural Language Toolkit. Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics. 2002. p. 63–70.

209. Regex Package [Internet]. Available from: https://pypi.python.org/pypi/regex

210. PubMed Central [Internet]. Available from: http://www.ncbi.nlm.nih.gov/pmc/

211. Tsuruoka Y, McNaught J, Ananiadou S. Normalizing biomedical terms by minimizing ambiguity and variability. BMC Bioinformatics. 2008;9 Suppl 3:S2.

212. HGNC [Internet]. Available from: http://www.genenames.org/

213. NCBI Gene: gene_info file [Internet]. [cited 2015 Jan 1]. Available from: ftp://ftp.ncbi.nih.gov/gene/DATA/gene_info.gz

214. UniProt: uniprot_sprot_human file [Internet]. Available from: ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/taxonomic_divisions/uniprot_sprot_human.dat.gz

215. HGNC: hgnc_complete_set file [Internet]. [cited 2015 Jan 1]. Available from: ftp://ftp.ebi.ac.uk/pub/databases/genenames/hgnc_complete_set.txt.

216. Hettne KM, van Mulligen EM, Schuemie MJ, Schijvenaars BJ, Kors JA. Rewriting and suppressing UMLS terms for improved biomedical term identification. J Biomed Semantics. 2010;1(1):5.

217. Schwartz AS, Hearst MA. A Simple Algorithm For Identifying Abbreviation Definitions in Biomedical Text. Pacific Symposium on Biocomputing. 2003. p. 451–62.

218. JSRE Tool [Internet]. Available from: https://hlt-nlp.fbk.eu/technologies/jsre

219. Chang C, Lin C. LIBSVM : A Library for Support Vector Machines. ACM Trans Intell Syst Technol. 2011;2:1–39.

220. LIBSVM [Internet]. Available from: http://www.csie.ntu.edu.tw/~cjlin/libsvm/

221. Tomanek K, Wermter J, Hahn U. A reappraisal of sentence and token splitting for life sciences documents. Stud Health Technol Inform. 2007;129:524–8.

222. Apache OpenNLP [Internet]. Available from: https://opennlp.apache.org/

223. Porter MF. An algorithm for suffix stripping. Program: electronic library and information systems. 1980. p. 130–7.

224. Liu H, Christiansen T, Baumgartner WA, Verspoor K. BioLemmatizer: a lemmatization tool for morphological processing of biomedical text. Journal of Biomedical Semantics. 2012. p. 3.

225. Bunescu RC, Mooney RJ. A shortest path dependency kernel for relation extraction. Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing. 2005. p. 724–31.

226. Becker KG, Barnes KC, Bright TJ, Wang SA. The genetic association database. Nat Genet. 2004;36(5):431–2.

227. Goh K-I, Cusick ME, Valle D, Childs B, Vidal M, Barabási A-L. The human disease network. Proc Natl Acad Sci U S A. 2007;104:8685–90.

228. PubAnnotation Project [Internet]. Available from: http://pubannotation.org/projects/DisGeNET

229. Liu H, Lussier Y a, Friedman C. A study of abbreviations in MEDLINE abstracts. Proceedings of AMAI symposium. 2002. p. 393–7.

230. Joshi M, Pedersen T, Maclin R. A Comparative Study of Support Vector Machines Applied to the Supervised Word Sense Disambiguation Problem in the Medical Domain. Proceedings of the 2nd Indian International Conference on Artificial Intelligence (IICAI05). 2005. p. 20.

231. Stevenson, Mark, Yikun Guo and RG. Acquiring Sense Tagged Examples using Relevance Feedback. Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1 Association for Computational Linguistics. 2008. p. 809–16.

232. Leroy G, Rindflesch TC. Effects of information and machine learning algorithms on word sense disambiguation with small datasets. Int J Med Inform. 2005;74(7-8):573–85.

233. Weeber M, Mork JG, Aronson a R. Developing a test collection for biomedical word sense disambiguation. Proceedings / AMIA . Annual Symposium AMIA Symposium. 2001. p. 746–50.

234. Alexopoulou D, Andreopoulos B, Dietze H, Doms A, Gandon F, Hakenberg J, et al. Biomedical word sense disambiguation with ontologies and metadata: automation meets accuracy. BMC Bioinformatics. 2009;10(1):28.

235. Le H-Q, Tran M-V, Dang TH, Ha Q-T, Collier N. Sieve-based coreference resolution enhances semi-supervised learning model for chemical-induced disease relation extraction. Database (Oxford). 2016;2016:1–14.

236. Koike A, Niwa Y, Takagi T. Automatic extraction of gene/protein biological functions from biomedical text. Bioinformatics. 2005;21(7):1227–36.

237. Khare R, Good BM, Leaman R, Su AI, Lu Z. Crowdsourcing in biomedicine: Challenges and opportunities. Brief Bioinform. 2016;17(1):23–32.

238. Semantic Knowledge Representation [Internet]. Available from: http://semrep.nlm.nih.gov/

239. AIMED Corpus [Internet]. Available from: ftp://ftp.cs.utexas.edu/pub/mooney/bio-data/

240. Bunescu RC, Mooney RJ. Subsequence kernels for relation extraction. Adv Neural Inf Process Syst. 2006;18:171.

241. Pinero J, Queralt-Rosinach N, Bravo À, Deu-Pons J, Bauer-Mehren A, Baron M, et al. DisGeNET: A discovery platform for the dynamical exploration of human diseases and their genes. Database. 2015;2015:1–17.

242. DisGeNET [Internet]. Available from: http://www.disgenet.org/

243. Gutiérrez-sacristán A, Grosdidier S, Valverde O, Torrens M, Bravo À, Piñero J, et al. PsyGeNET : a knowledge platform on psychiatric disorders and their genes. 2015;11–3.

244. PsyGeNET [Internet]. Available from: http://www.psygenet.org/

245. Vincze V, Szarvas G, Farkas R, Móra G, Csirik J. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. BMC Bioinformatics. 2008;9(Suppl 11):S9.

246. Kilicoglu H, Bergler S. Recognizing speculative language in biomedical research articles: a linguistically motivated perspective. BMC Bioinformatics. 2008;9(Suppl 11):S10.

247. Nawaz R, Thompson P, Ananiadou S. Negated bio-events: analysis and identification. BMC Bioinformatics. 2013;14:14.

248. Martin EPG, Bremer EG, Guerin M-C, DeSesa C, Jouve O. Analysis of protein/protein interactions through biomedical literature: Text mining of abstracts vs. text mining of full text articles. Knowledge Exploration in Life Science Informatics. Springer B. 2004. p. 96–108.

249. SEKI K, MOSTAFA J. Discovering implicit associations between genes and hereditary diseases. Pacific Symposium on Biocomputing. 2007. p. 316–27.

250. Kou Z, Cohen WW, Murphy RF. A stacked graphical model for associating sub-images with sub-captions. Pacific Symp Biocomput. 2007;257–68.

251. Jun Xu, Yonghui Wu, Yaoyun Zhang, Jingqi Wang, Ruiling Liu, Qiang Wei and HX. UTH-CCB@BioCreative V CDR Task: Identifying Chemical-induced Disease Relations in Biomedical Text. Proceedings of the Fifth BioCreative Challenge Evaluation Workshop. 2015. p. 254–9.